DOCUMENT RESUME

ED 380 278	SE 055 878

AUTHOR	Cipra, Barry
TITLE	What's Happening in the Mathematical Sciences, 1993-1994.
INSTITUTION	American Mathematical Society, Providence, R.I.
REPORT NO	ISBN-0-8218-8998-2; ISBN-0-8218-8999-0; ISSN-1065-9358
PUB DATE	93
NOTE	109p.; Published annually, starting in 1993.
AVAILABLE FROM	American Mathematical Society, P.O. Box 5904, Boston, MA 02206-5904 (order no. for volume 1: HAPPENING/1wh, $7; order no. for volume 2: HAPPENING/2WH, $8).
PUB TYPE	Collected Works - Serials (022)
JOURNAL CIT	What's Happening in the Mathematical Sciences; v1-2 1993-1994
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Algorithms; Biology; Classification; Coding; Computers; Crystallography; Environmental Education; Geometry; Higher Education; *Innovation; *Mathematical Applications; *Mathematics Instruction; Prime Numbers; Proof (Mathematics); Science Education; Secondary Education
IDENTIFIERS	*Mathematical Sciences; Medical Technology

ABSTRACT

This document consists of the first two volumes of a new annual serial devoted to surveying some of the important developments in the mathematical sciences in the previous year or so. Mathematics is constantly growing and changing, reaching out to other areas of science and helping to solve some of the major problems facing society. Volumes 1 and 2 survey some of the important developments in the mathematical sciences over the past year or so. The contents of volume 1 are: (1) "Equations Come to Life in Mathematical Biology"; (2) "New Computer Insights from 'Transparent' Proofs"; (3) "You Can't Always Hear the Shape of a Drum"; (4) "Environmentally Sound Mathematics"; (5) "Disproving the Obvious in Higher Dimensions"; (6) "Collaboration Closes in on Closed Geodesics"; (7) "Crystal Clear Computations"; (8) "Camp Geometry"; (9) "Number Theorists Uncover a Slew of Prime Impostors"; and (10) "Map-Coloring Theorists Look at New Worlds." The contents of volume 2 are: (1) "A Truly Remarkable Proof" (Fermat's Last Theorem); (2) "From Knot to Unknot"; (3) "New Wave Mathematics"; (4) "Mathematical Insights for Medical Imaging"; (5) "Parlez-vous Wavelets?" (6) "Random Algorithms Leave Little to Chance"; (7) "Soap Solution"; (8) "Straightening Out Nonlinear Codes"; (9) "Quite Easily Done"; and (10) "(Vector) Field of Dreams." (MKR)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

What's Happening in the Mathematical Sciences

Volume 1 • 1993

ED 380 278

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

PAUL G.
CHAMBERS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

American Mathematical Society

Introduction

Welcome to the inaugural issue of *What's Happening in the Mathematical Sciences*! To be published annually, *What's Happening* surveys some of the important developments in the mathematical sciences over the past year or so. Mathematics is constantly growing and changing, reaching out to other areas of science and helping to solve some of the major problems facing society. Here you can read about the development of a mathematical model of the human heart, the solution to a longstanding mathematical problem about the way a drum's shape affects its sound, and the contributions mathematics is making to the solution of environmental problems.

What's Happening in the Mathematical Sciences aims to inform the general public about the beauty and power of mathematics. The American Mathematical Society is pleased to present this new publication. We hope you enjoy it!

Samuel M. Rankin, III
AMS Associate Executive Director
and Director of Publications

Cover Illustration. A group of scientists at Los Alamos National Laboratory have developed a mathematical model of ocean dynamics for massively parallel computers that they hope will improve understanding of the role of oceans in global climate change. The colors in this computer-generated picture indicate sea surface temperature from cold (blue) to warm (red). Figure courtesy of Richard Smith, John Dukowicz, and Robert Malone at Los Alamos National Laboratory.

Contents

Equations Come to Life in Mathematical Biology 3

Mathematicians are working with biologists to delve into some of the most challenging problems in biology today, from understanding the human immune system to "computing" the human heart.

New Computer Insights from "Transparent" Proofs 7

Can a computer be trusted when it produces a proof so long and complicated that no human can check the details? Theorists have cooked up a new way to tell whether or not a computer proof is right.

You Can't Always Hear the Shape of a Drum 13

Can you hear the shape of a drum? is a famous problem that asks if two drums that look different can make the same sound. After decades of head-scratching, mathematicians have come up with the answer.

Environmentally Sound Mathematics 17

Mathematicians have been teaming up with scientists to work on solving environmental problems, from ocean modeling to dealing with hazardous waste.

Disproving the Obvious in Higher Dimensions 21

Intuition about our three-dimensional world can be surprisingly misleading when it comes to higher dimensions, as two recent results in geometry show.

Collaboration Closes in on Closed Geodesics 27

An unusual blend of differential geometry and dynamical systems has led to an important theoretical result about the number of closed "geodesic" curves on distorted spheres.

Crystal Clear Computations 31

Growing crystals—on a computer? Mathematicians are helping materials scientists to better understand the nature of crystals, while picking up some challenging mathematical problems along the way.

Camp Geometry 35

A group of talented and inquisitive undergraduates "camped out" last summer at the Geometry Center. Using sophisticated computer graphics and their own imaginations, they came up with some fascinating mathematics.

Number Theorists Uncover a Slew of Prime Impostors 39

Strange as it may sound, there are composite numbers that "masquerade" as primes. A group of mathematicians trying to hunt down these prime impostors ended up proving there are infinitely many of them.

Map-Coloring Theorists Look at New Worlds 43

How many colors are needed to distinguish neighboring countries on a map? The famous Four Color Theorem notwithstanding, this is a challenging problem in graph theory—especially when your maps aren't flat.

ISBN 0-8218-8999-0, ISSN 1065-9358

© 1993 by the American Mathematical Society. All rights reserved.

Permission is granted to make and distribute verbatim copies of this publication or of individual items from this publication provided the copyright notice and this permission notice are preserved on all copies.

Permission is granted to copy and distribute modified versions of this publication or of individual items from this publication under the conditions for verbatim copying, provided that the entire resulting derived work is distributed under the terms of a permission notice identical to this one.

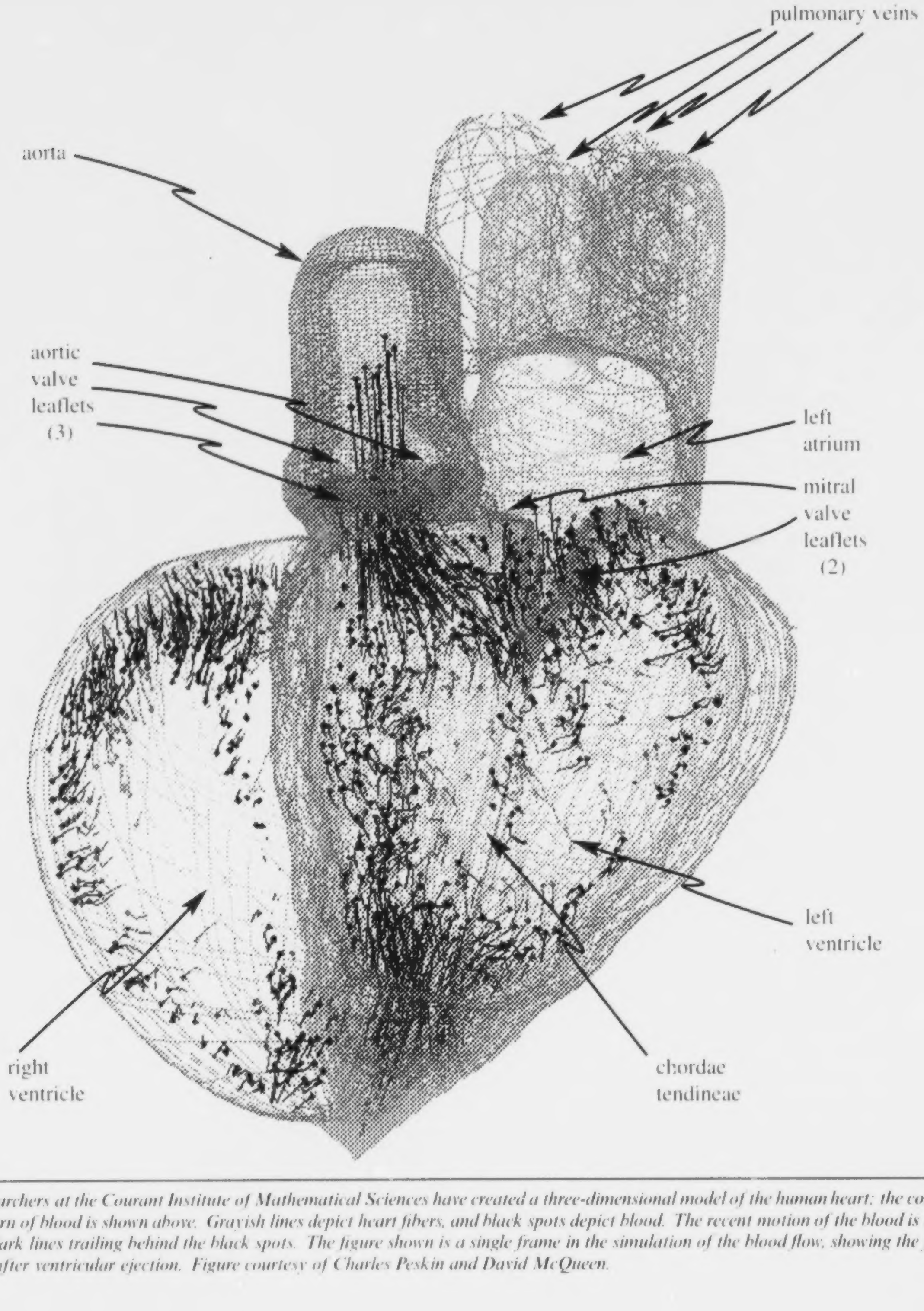
1991 *Mathematics Subject Classification*: Primary 00A06.

Printed in the United States of America.

This publication has been typeset using the TEX typesetting system running on a Solbourne 5/502 Unix computer. Halftones were created from original photographs with Adobe Photoshop and illustrations were redrawn using Adobe Illustrator on Macintosh Quadra and Macintosh IIci computers. PostScript code was generated using `dvips` by Radical Eye Software.

Typeset on an Agfa/Compugraphic 9600 laser imagesetter at the American Mathematical Society. Printed at E. A. Johnson, East Providence, RI, on recycled paper.

♻


pulmonary veins
aorta
aortic valve leaflets (3)
left atrium
mitral valve leaflets (2)
left ventricle
right ventricle
chordae tendineae

Researchers at the Courant Institute of Mathematical Sciences have created a three-dimensional model of the human heart: the computed flow pattern of blood is shown above. Grayish lines depict heart fibers, and black spots depict blood. The recent motion of the blood is indicated by the dark lines trailing behind the black spots. The figure shown is a single frame in the simulation of the blood flow, showing the flow pattern just after ventricular ejection. Figure courtesy of Charles Peskin and David McQueen.

Equations Come to Life in Mathematical Biology

The Nile crocodile and the Egyptian plover have a fascinating relationship. The croc, ordinarily a surly saurian, will sit placidly on the muddy river bank, mouth wide open, while the bird hops from tooth to tooth scarfing leeches and other tasty morsels. *Crocodylus niloticus* enjoys a thorough oral prophylaxis; *Pluvianus egyptius* gets a meal.

The technical term is *symbiosis*.

Something like that is evolving between biologists and mathematicians. Biology has a host of problems that call out for mathematical analysis, from the folding of proteins inside an individual cell to the complex food webs on the ocean floor. Mathematics, for its part, provides a quantitative framework that can bring order to the organic chaos of nature and point toward new directions for research. Mathematics has brought new insights into biology; biology has inspired new mathematical results.

Which you regard as the bird and which the crocodile is a matter, shall we say, of taste.

Mathematics and biology are not exactly newcomers to each other. Mathematical methods have long been used in population studies, epidemiology, genetics, and physiology, to name a few. And biological problems have spurred the creation of many mathematical techniques, including, arguably, the entire field of statistics.

What's new is the depth of detail that mathematical models are now striving for—and the attendant depth of theory required. The problems being tackled today call for closer cooperation than ever before between mathematics and biology. Increasingly, mathematicians are getting in on the ground floor of biological research, working directly with biologists to help tease out the mathematical structure in phenomena ranging from the undulating motion of fish to the beating of the human heart.

"The field's very different now than it was thirty years ago," says Alan Perelson, a mathematical biologist at Los Alamos National Laboratory and president of the Society for Mathematical Biology. "Early mathematical biology was really mathematics with a little inspiration from biology." There was little real communication between the fields. But the current generation of mathematical biologists, Perelson says, consists of researchers "who've been driven by the biology, who look at the details, talk to experimentalists, and generate models that are attempting to answer questions of interest to experimentalists."

Perelson's own work has been in theoretical immunology. He and others in the field are trying to develop mathematical models for the sequence of events that begins when, say, you step on a rusty nail, from the first antigenic signals presented by the invading bacteria, to the final chemotactic processes that close the wound either cleanly or with a lasting scar. It's not just a matter of programming a computer to do a bunch of calculations. Researchers first have to identify the crucial biological aspects of the process and then find the appropriate mathematical equations that describe them. Developing such a thorough understanding,

Biology has a host of problems that call out for mathematical analysis, from the folding of proteins inside an individual cell to the complex food webs on the ocean floor.

Perelson says, is the "grand goal" of theoretical immunology, but that goal is still a long way off. "We are really at the very beginning."

One reason for that is the daunting complexity of the immune system. The body's response to the variety of pathogens it encounters is, among other things, a pattern-recognition problem: The body must somehow identify an invading virus or bacteria based on the invader's distinctive pattern of chemical clues. The immune system's ability to do this, researchers believe, depends on the diversity of its receptors. "To do pattern recognition [for the immune system] seems to require on the order of ten million different types of receptors," Perelson explains. "So to understand in a profound sense how the system operates — to recognize pathogens and respond — one really has to deal with systems of enormous complexity." Mathematics enters the picture as a tool for modeling not only the individual receptors, but also the overarching structure that enables them to act in concert.

Nancy Kopell. (Photo by Janet Coleman.)

The emergence of organized behavior from a collection of individual entities is not unique to the immune system; it is a hallmark of living systems. A central problem in biology is to deduce how properties of a system at one level of organization produce behavior at higher levels — for example, how does the electrical activity in the nervous system of a centipede organize itself into the correct patterns to make the critter's legs move in a coordinated fashion?

Nancy Kopell, a mathematician at Boston University, likens this problem to the task of figuring out how a television works knowing only the properties of transistors. She sees the modeling of "emergent behavior" as a central concern for mathematical biology. "There are many questions in biology involving the behavior of systems in which what you can measure easily... is the behavior of some of the components of the system," Kopell says. "What you can't easily, or sometimes not at all, get from direct measurements is what's going to happen when you hook all these things up. For that you really need some kind of theory."

Kopell and her mathematical colleague Bard Ermentrout of the University of Pittsburgh have collaborated with biologists to study the rhythmic neuronal patterns that give rise to swimming in an eel-like fish called a lamprey. Researchers had known for some time that the electrical activity in the lamprey spinal cord could be represented mathematically as a "chain of oscillators" — something like a set of pendulums hooked together by springs, but with quite different mathematical properties. Kopell and Ermentrout formulated a new mathematical model based on a deeper analysis of how the oscillators are hooked together. Their model produced predictions which could be verified by experimentalists, and it provided new insight into how the electrical activity organizes itself to produce the swimming motion in lampreys. The model also helped point out new directions for biological research. And as new data from new experiments is found, Kopell and Ermentrout continue to refine their mathematics to better reflect the biology.

Computer simulation figures prominently in many of the modeling efforts in mathematical biology. Indeed, revolutions in both hardware and software have been crucial to advances across the board. The Human Genome Project, with its ambitious goal of mapping the roughly three billion base pairs that constitute human DNA, would be inconceivable without machines and mathematical algorithms for dealing with vast amounts of data. (It's not just a question of storing three billion pieces of information; it's a question of *analyzing* that data.) Likewise, mathematics is at the heart of much of medical imaging, including CAT scans, nuclear magnetic resonance, and positron emission tomography. These techniques are made possible by machines that carry out mathematical manipulations of the data that pour into them.

Mathematics is at the heart of much of medical imaging, including CAT scans, nuclear magnetic resonance, and positron emission tomography.

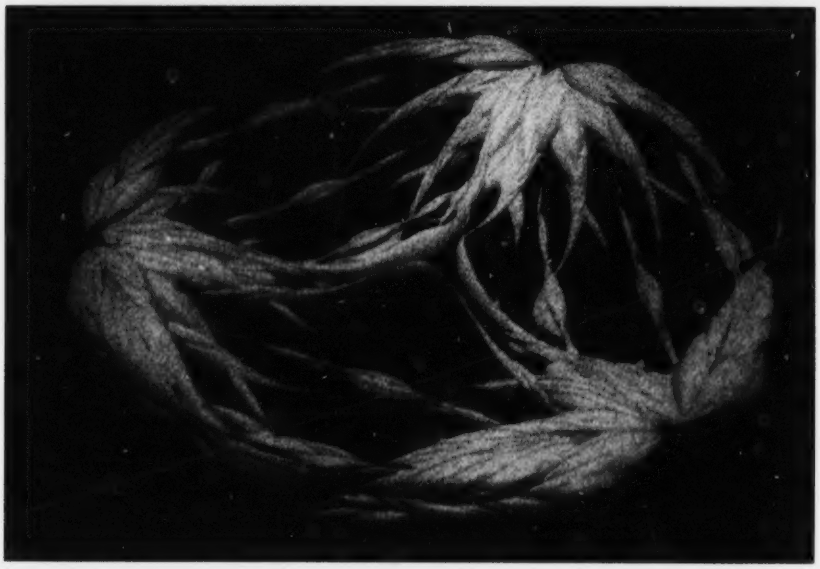
One notable example of the use of mathematics and computer simulation in physiology is the work of Charles Peskin and colleagues at the Courant Institute of Mathematical Sciences at New York University. They are in the process of building a realistic three-dimensional mathematical model of the human heart. The model, they hope, will give researchers insight into the functioning—and malfunctioning—of real hearts and lead to improved designs for artificial valves and other replacement parts.

"It's a very large effort, and it's still going on," Peskin notes. The model is nearly complete anatomically, but "we're still working on getting the physiology right," he adds. That means figuring out the appropriate elasticities of the parts, how fast they should contract, and how fast they should relax, and then fine-tuning the equations to reflect these physiological attributes.

The geometry of the heart is also a crucial part of the model. Conceptually, the Courant heart consists of hundreds of closed curves representing muscle fibers. "In effect the [model] heart is constructed out of a very large array of rubber bands," Peskin explains. Mathematically, the curves are represented by a string of discrete points, with specified spring-like elasticity between each pair of consecutive points.

A computer keeps track of all these points—on the order of a million of them—and immerses them in a computer-simulated bath of blood. Then the real calculation begins: The numerical heart starts to beat.

The mathematics of the calculation can be described by something that sounds like the title of a 1950s Japanese monster movie: Hookes's Law Meets the Navier-Stokes Equation. Hookes's Law is the force-displacement relation for springs, familiar from high-school physics; a fancier, nonlinear version of it is used to



Detail from the three-dimensional Courant heart, showing the three leaflets of the model aortic valve in its closed position. The fiber architecture of the valve has a fractal structure which has been predicted here by solving an equation for the mechanical equilibrium of the fibers under a pressure load. (Illustration created at the Pittsburgh Supercomputing Center.)

With the concurrent revolutions in both biology and applied and computational mathematics, [Peskin] says, "the kinds of problems that we can realistically hope to do are expanding tremendously."

David McQueen and Charles Peskin. (Photo reprinted with permission of Projects in Scientific Computing, *Pittsburgh Supercomputing Center.)*

model the heart's muscle fibers. The Navier-Stokes equation, while less familiar, is even more universal: It describes fluid flow of virtually any kind, from blood pumping through the heart to global circulation patterns of the earth's atmosphere.

These are the basic mathematical ingredients that determine the complex motions of the heart and the blood moving through it. Unfortunately, you can't sit down with pencil and paper and solve the equations precisely—the solutions are only approachable by computer approximation. And that turns out to be a formidable task, even for a supercomputer. Solving fluid flow problems is always computationally demanding, but the heart model presents a particular challenge: Unlike flow down a pipe or past a spinning turbine, where the boundary of the fluid is fixed or moving in a prescribed manner, the motion of the heart wall is among the unknowns that must be solved for.

"You not only don't know the boundary velocity, you don't know where the boundary is," notes David McQueen, a mechanical engineer who has collaborated with Peskin for the past fifteen years. "Your traditional engineering approach is going to be hard pressed to solve this problem."

Instead, Peskin has developed mathematical techniques for what he calls "immersed boundary" problems. "The beauty of this method is that it allows you to do computing in situations where you don't know the boundary motions in advance," says McQueen. Modeling heartbeats is not the only application. "The technique is generally useful in biofluid dynamics," Peskin says, "and it has already been applied to a wide variety of problems such as platelet aggregation during blood clotting, aquatic animal locomotion, and wave propagation in the inner ear." Peskin anticipates future applications in the study of flow in collapsible tubes such as thin-walled blood vessels, flow in renal (kidney) tubules, and the flight of birds and bats. There are even nonbiological possibilities, such as the design of aerodynamically efficient sails and parachutes.

The current heart model is a step up from a two-dimensional heart that Peskin began developing in the early 1970s. Paradoxically, Peskin notes, the 2-D heart is still, in some ways, more realistic than the 3-D model. That's mainly because the extra effort of computing in three dimensions has forced the modelers—for now—to use a simpler muscle model. Further advances in both theory and hardware will undoubtedly bring the 3-D model up to speed, but the 2-D heart is likely to continue being used for experimental computations. "What we'd really like is to use the 2-D model as a way of getting rough results, and then perhaps do a few 3-D computations to verify those findings," McQueen says.

Indeed, the 2-D model has already proved useful in artificial heart valve design. By experimenting with the shape of a prosthetic mitral valve (the gate between the left atrium and ventricle), McQueen and Peskin found a design that simultaneously increased the flow velocity near the valve and reduced the pressure drop across it—two features that are prized in artificial valves. While not yet in clinical use, the design has been patented and licensed.

The 3-D model has not yet had any such applications, but those are likely to come as the model becomes more physiologically realistic and as the computing demands get more manageable. (Currently a single beat takes upwards of fifty hours of supercomputer time.) Peskin sees the heart model, and other models in the future, as important experimental tools. With the concurrent revolutions in both biology and applied and computational mathematics, he says, "the kinds of problems that we can realistically hope to do are expanding tremendously."

New Computer Insights from "Transparent" Proofs

Mathematicians are professional skeptics. When told of a new result, their first response is, Where's the proof? Even when shown a proof, they're not completely convinced it's correct until they check every last line.

This professional skepticism isn't limited to traditional mathematical proofs. It extends to results produced by computers as well. Today's lightning-fast, high-tech adding machines take the labor out of long, laborious calculations, making it possible to carry out computations that could never be done by hand. But they leave behind the lingering question, Did the computer do its job correctly?

A sequence of recent breakthroughs in theoretical computer science may put that question to rest. Researchers have found some unexpected new ways by which computers can prove "beyond a shadow of a doubt" that the results they provide are indeed reliable. Moreover, these developments are giving theorists new insights into some of the hardest problems of computer science.

Guaranteeing the reliability of computer results is obviously of concern to more than mathematicians. But by thinking of computations themselves as proofs that certain inputs produce certain outputs, theoretical computer scientists are able to view anything a computer does in logical mathematical terms. Moreover, the computational aspects of many problems can be recast as purely mathematical questions in areas such as graph theory or elementary, first-order logic. The abstract language of mathematics helps clarify the essential issues, which might otherwise be lost among the details of individual applications.

Some computations are easy enough to check. For example, researchers often need to know if there is a path that travels along the edges of a graph, visiting each vertex once and only once—what graph theorists call a "Hamiltonian cycle." (This kind of problem crops up in applications such as designing efficient telecommunications networks.) If a computer says there is a Hamiltonian cycle, it can prove it simply by pointing out the path (as done with dark lines in Figure 1a). But when it says there is *no* such path for the graph in Figure 1b, how can you be sure it didn't overlook one—or, worse, that your computer saw one but chose not to tell you?

The computer can, of course, produce a proof by trying all possible routes around the graph and showing that none is a Hamiltonian cycle. That's not an unreasonable thing to do for Figure 1b. But the number of possible routes grows so quickly with the number of vertices that this straightforward approach soon becomes unwieldy. For graphs that typically occur in telecommunications network problems, for example, this kind of proof would take inconceivably long even on the fastest conceivable supercomputer. That defeats the purpose of having a fast machine. Worse, one is still left with the task of checking that all the computations were done correctly.

The problem is, errors in a proof don't always, or even usually, call attention to themselves—and all it takes to invalidate an entire proof is one mistake, as minor as a misplaced minus sign. "Mathematical proofs are very fragile," says László Babai, a theoretical computer scientist at the University of Chicago. Like a string

Today's lightning-fast, high-tech adding machines take the labor out of long, laborious calculations. But they leave behind the lingering question, Did the computer do its job correctly?

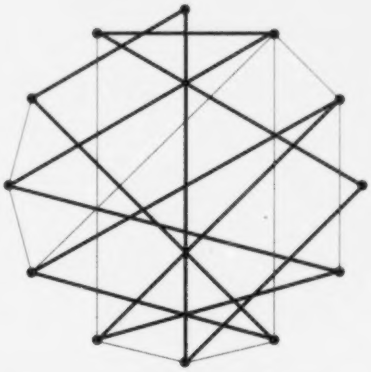


Figure 1a. *The dark edges "prove" that this graph has a Hamiltonian cycle.*

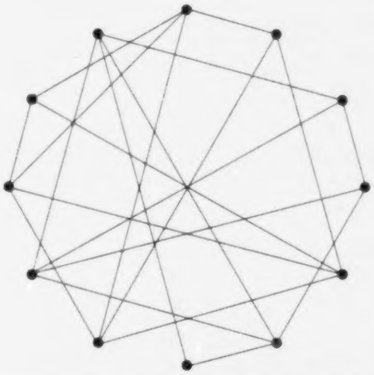


Figure 1b. *This figure does not have a Hamiltonian cycle—or does it?*

of pearls, if the strand breaks anywhere, the whole necklace winds up scattered on the floor—only with a proof, you might not notice till you've left the party.

However, help is on the way—sort of. Over the last decade, Babai and others have developed techniques by which even unreliable computers can, in principle, provide overwhelming evidence that their calculations are correct. Researchers have recently shown it's possible for a computer to reformulate an ordinary computational proof in such a way that the correctness of the original proof can be guaranteed—with near certainty—by merely "spot reading" the transformed version at a relatively small number of randomly chosen places.

That may not satisfy mathematicians, for whom being "nearly certain" is worth about as much as a basketball player's last-second, game-winning three-pointer that "nearly went in." The current techniques are also far from practical—it is unlikely the spot-checking techniques will ever be used directly to test the veracity of computers' output. However, the theory has paid off handsomely in other ways, mainly by giving researchers new insights in the theory of computational complexity—the study of how hard a computer has to work to arrive at an answer.

In particular, researchers have discovered an astounding connection with a seemingly unrelated issue in complexity theory: the question of whether there can be "easy" ways to approximate the solutions to computational problems in a class known as NP—problems that are thought to be intrinsically "hard" to solve exactly (see box on next page). Surprisingly, the existence of spot-checkable proofs turns out to preclude the existence of "easy" approximation algorithms for a substantial subset of the problems in NP—unless there are easy exact algorithms for the whole class NP, a prospect few in the computer science community believe to be the case.

The new results stem from work on "interactive proofs," a notion that was introduced in the mid-1980s by Shafi Goldwasser and Silvio Micali at MIT and Charles Rackoff at the University of Toronto. An interactive proof is a lot like a police interrogation. A "verifier" (the detective) asks a "prover" (the suspect)

Lance Fortnow. (Photo by Matthew Gilson, University of Chicago.)

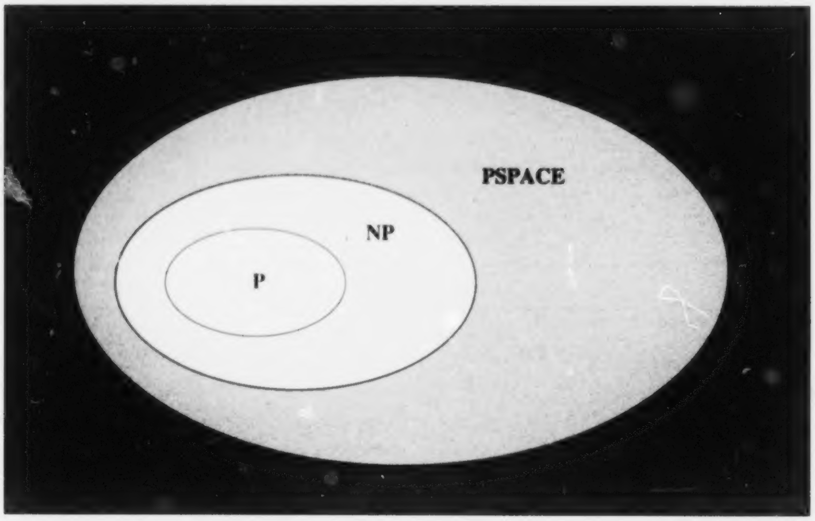


Figure 2. *Computational problems for which efficient algorithms exist constitute the smallest in a hierarchy of "complexity" classes.*

A Complexity Primer

The difference between "easy" and "hard" is at the heart of theoretical computer science. In essence, a computational problem is "easy" if the number of steps required to solve it is bounded by some power of the size of the problem. For example, multiplication of two N-digit numbers is "easy" because it takes at most N^2 single-digit multiplications and additions. Such problems form a class which computer scientists call P, for *polynomial time* problems (see Figure 2).

The class P contains a great many problems, including such significant computational tasks as linear programming. But a great many more problems seem to lie just beyond it, problems whose computational demands apparently grow exponentially with size. These are the problems that complexity theory calls "hard."

In particular, there's the tantalizing class of decision problems known as NP. (Decision problems are problems for which a simple Yes/No answer is sought. Technically P also consists strictly of decision problems, but when answers are easy to come by, the technicality is unimportant.) The problems in NP (which stands for *nondeterministic polynomial time*) have a curious dual nature: The amount of computation required to arrive at the Yes/No answer may grow exponentially with the size of the problem, but, at least when the answer is Yes, an "inspired guess" can reduce the amount of computation down to a simple, polynomial-time verification.

The Hamiltonian cycle problem is one example. The number of possible paths in a graph grows exponentially with the number of vertices and edges, but if a cycle exists, then all that computation is rendered unnecessary if someone simply tells you which path works, and you simply check it out. That can happen if the given graph was created by first drawing a Hamiltonian cycle and then disguising it with additional edges. In a sense, a problem in NP is a little like a riddle—only in NP, the riddle may have no answer.

But why, one may ask, is it necessary to check all possible paths in order to determine whether or not a graph has a Hamiltonian cycle? Is it not possible that some other method could arrive at the answer without going through an exponential number of cases? Is it not possible that the Hamiltonian cycle problem in fact is "easy"—a member of P—and it just looks hard because no one has found a polynomial-time algorithm for it yet?

Good question. In fact, that's the core conundrum of computer science. Nobody knows if there really are "hard" problems in the class NP; the classes P and NP may be one and the same. The lack of an easy algorithm for problems like the Hamiltonian cycle problem may be due not to the nonexistence of such an algorithm, but to the limits of mathematicians' ingenuity. It's within the realm of possibility that someone could find an easy algorithm for these hard problems.

It's happened before. Linear programming was long classified as a hard problem because the simplex method was known to suffer the kind of exponential computational growth that's characteristic of NP problems. But then, in 1979, Leonid Khachian of the USSR Academy of Sciences discovered a polynomial-time—i.e., "easy"—algorithm for solving linear programming problems. Hence a problem that had been thought of as hard turned out not to be so hard after all.

It could happen again. But if it happens for the Hamiltonian cycle problem, or any of several thousand other NP problems, there will be a huge fallout. That's because over the last two decades researchers have established a web of relationships among the problems in NP. Specifically, there is a subclass of NP, the so-called "NP-complete" problems, which have the property that any algorithm for solving any one of them can be translated into an algorithm for solving any other problem in the entire class NP.

The Hamiltonian cycle problem is NP-complete. So is the well-known Traveling Salesman Problem. So are many other problems in graph theory, combinatorics, and logic. If anyone ever finds a polynomial-time algorithm for any one of these, the distinction between P and NP will collapse; only P will remain.

Most theorists consider that unlikely. Only a few die-hard optimists believe that all problems (at least the ones in NP) are easy; the smart money says that NP really does contain hard problems. But so far neither side is anywhere close to collecting on the bet.

The theory has paid off handsomely in other ways, mainly by giving researchers new insights in the theory of computational complexity—the study of how hard a computer has to work to arrive at an answer.

László Babai. (Photo by Matthew Gilson, University of Chicago.)

Transparent proofs are unaffected by minor copying errors or other computer glitches. In essence, a transparent proof replaces the original proof's single strand of logic with a highly redundant cable.

a series of questions about the problem the prover claims to have solved. The questions are designed to expose any lie (or mistake) in the prover's answers. In effect, interactive proofs are the embodiment of Walter Scott's familiar warning, "O, what a tangled web we weave, When first we practice to deceive."

In order to prevent a "mastermind" prover from anticipating the verifier's questions and concocting a consistent "alibi" to support its original claim, the questions are chosen partly at random. Because of this, there's a chance that an interactive proof will occasionally put its stamp of approval on an incorrect result, just as a lazy student can occasionally guess his or her way to a perfect score on a True/False test. But the chance of that happening can be made as small as you like by simply asking more questions.

Interactive proof turns out to be a powerful tool. In 1989, researchers established that interactive proofs can be used to verify solutions for a large class of problems called PSPACE. Then in early 1990, Babai, Lance Fortnow, and Carsten Lund at the University of Chicago proved what initially looked like an innocent generalization: They showed that problems in an even larger class called NEXP could be verified by a "multi-prover" variant of interactive proofs.

A multi-prover interactive proof can be thought of as an interrogation of *two* suspects who have been separated for questioning. The intuitive idea is that it's easier to get two suspects to contradict one another than it is to get a single suspect to trip over his or her own story. The Chicago theorists made that intuition precise. In demonstrating the exact power of multi-prover interactive proofs, they paved the way for what came next: "transparent" proofs.

The notion of transparent proof was introduced by Babai and Fortnow in joint work with Leonid Levin at Boston University and Mario Szegedy at the University of Chicago. In essence, they found that the question-and-answer format of an interactive proof is unnecessary; instead, all that's needed is to have the prover rewrite its proof as a kind of legal deposition — but one that's "easy to see through" if the prover tries to lie. This "transparent" proof is a long, rambling retelling of the original proof, couched in a kind of computer-science legalese, consisting of purportedly true statements which can be checked against each other for accuracy and consistency.

The key is that the correctness of a transparent proof can be checked without reading the whole proof, or even very much of it. Any error in the original proof, no matter how small, is magnified and spread throughout the transformed version so that it becomes glaringly obvious. By "spot checking" a relatively small number of randomly chosen passages of the transparent proof, the verifier — who can now be thought of as a judge — either finds a definite mistake or concludes, with very high confidence, that the original proof was correct.

This also means that transparent proofs are unaffected by minor copying errors or other computer glitches. In essence, the transparent proof replaces the original proof's single strand of logic with a highly redundant cable.

"You take a proof, which is fragile, and you turn it into a very sturdy thing," says Babai. In other words, if a transparent proof isn't riddled with errors, then the original proof is probably actually okay.

But how much spot checking is needed to be sure? Babai and coworkers showed that if the original proof was N bits long (remember, everything a computer does boils down to a string of ones and zeros), then the transparent proof could be written in such a way that the number of spot checks required to verify the correctness of the original is proportional to a power of log N, such as $100(\log N)^2$. That difference can be appreciated by comparing $N = 1{,}000{,}000$ to $\log N = 6$. What's most important is that any multiple of any power of log N is eventually an insignificant fraction of N, so for very long proofs the amount of spot checking will be relatively small.

This was taken a step further by Shmuel Safra at Stanford University and the IBM Research Center at Almaden and Sanjeev Arora, a graduate student at the University of California at Berkeley. Safra and Arora found a way to write transparent proofs that could be checked by looking at only about log log N bits. Taken literally (using logs base 10), that implies that an original proof of length ten billion (10^{10}) could be checked by looking at a single bit of the transparent version!

But Safra and Arora weren't just out to reduce the spot-checking requirement of transparent proofs. They were after bigger game: an application of the new theory to an old and very important problem in computer science.

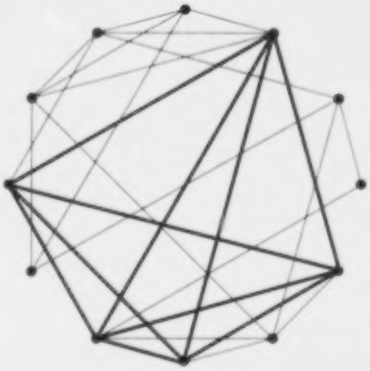


Figure 3. *The five vertices connected by the dark edges form a "clique" because there is an edge between each pair of points.*

Shortly after the introduction of transparent proofs, Safra, Szegedy, and Goldwasser, together with Uri Feige and László Lovász at Princeton University, found an unexpected connection between interactive proofs and a particular problem in graph theory: that of approximating the largest "clique" in a graph of N vertices. A clique is simply a subset of vertices that are pairwise adjacent (meaning that there's an edge connecting each pair of vertices) (see Figure 3). The problem of determining the *exact* size of the largest clique in a graph is known to be NP-hard—that is, any efficient algorithm for solving this one problem would translate easily into efficient algorithms for solving any problem in the class NP.

What the five researchers showed was that the problem of *approximating* the size of the largest clique is "very nearly" NP-hard. In other words, if the size of the largest clique can be approximated—even poorly—by an efficient algorithm, then any problem in NP can be solved by algorithms that are "very nearly" efficient.

Safra and Arora removed those adverbs. Their refinement of transparent proofs implies that if the largest-clique problem could be solved approximately by an efficient algorithm, then there would be truly efficient algorithms for all problems in NP. In the jargon of computer science, NP would equal P.

That implication was soon extended from the largest-clique problem to a host of other approximation problems by Arora and fellow graduate student Madhu Sudan at Berkeley, Rajeev Motwani at Stanford, and Lund and Szegedy, both now at AT&T Bell Laboratories. They did so by pushing transparent proofs to an extreme: In their approach, all transparent proofs can be verified with the same number of spot checks no matter how long the original proofs are.

The only thing better would be a transparent proof you didn't have to read at all!

The string of breakthroughs in this area of computational complexity came in rapid succession—as befits a subject concerned with speed and efficiency. The implications for computer science—both theoretical and practical—are yet to be sorted out. Researchers want to know whether transparent proofs can be streamlined to more manageable lengths. They also are finding more problems that look "hard" to approximate. Finally, computer scientists continue to ponder what these results say about the class NP—in particular, are these seemingly hard problems really all that hard to solve?

The answer to *that* question is sure to have everyone looking carefully at the proof.

The string of breakthroughs in this area of computational complexity came in rapid succession—as befits a subject concerned with speed and efficiency.

WASH. U. MATH
WASH. U. MATH

David Webb and Carolyn Gordon with paper models of a pair of "sound-alike" drums. (Photo courtesy of Washington University in St. Louis.)

BEST COPY AVAILABLE

You Can't Always Hear the Shape of a Drum

Much of what scientists know of the natural world comes not from direct observation, but by means of indirect measurements. Astronomers, for example, cannot sample the stuff of stars; instead they infer stars' composition by analyzing spectrographic images. Likewise, geophysicists construct a picture of the planet's interior from seismic studies, not from journeys to the center of the earth. X-rays, CAT scans, and other medical imaging techniques are also indirect ways of seeing inside the body. Even your family doctor prefers the stethoscope to the scalpel.

Mathematically, the job of reconstructing an object out of measurements of certain "observable" properties is known as an inverse problem. (A "direct" problem is to deduce observable properties from explicit knowledge of an object.) There are many important questions about inverse problems that mathematicians and others have worked to resolve, such as how many measurements are necessary to get an answer and how much accuracy is required. But underlying these questions is a deeper mathematical question: Even if you can take infinitely many measurements with infinite precision, can you be sure of your conclusions? Or to put it differently, can two different objects look alike in every measurable way?

It might seem the answer to this question should be obviously yes. But it's a lot more difficult than that — and that's where mathematical theory steps in. In 1966, the Polish-American mathematician Mark Kac zeroed in on a particular inverse problem. Can one, Kac queried, hear the shape of a drum?

That may seem like a strange question at first, but it's no stranger than asking if one can "see" the chemistry of a star or "hear" the interior of the earth. Moreover, Kac's question has a precise mathematical meaning. The problem it poses had been a challenge for more than fifty years at the time of Kac's lecture, and it continued to stymie researchers for another three decades. Then, finally, in the spring of 1991, three mathematicians — Carolyn Gordon and David Webb at Washington University in St. Louis, and Scott Wolpert at the University of Maryland — came up with the answer: a resounding No.

Gordon, Webb, and Wolpert found a pair of distinct geometric shapes in the plane which, when thought of as mathematical drums, resonate at the exact same frequencies. In other words, if your goal is to deduce the shape of a drum merely from the sounds it makes, these two drums provide an example where that goal cannot be achieved: You can't decide which drum you're listening to, because they both sound the same.

That's more than musically important, according to Dennis DeTurck of the University of Pennsylvania, an expert on "isospectral geometry," as the mathematical theory of such inverse problems is called. It points out there are subtle mathematical questions involved whenever scientists attempt to reconstruct reality from a set of data. The fact that even in a relatively simple mathematical setting there is not always just one conclusion that can be reached from a complete set of measurements is, to put it mildly, unsettling.

Remarkably, the final proof that the pair of sound-alike drums actually do sound

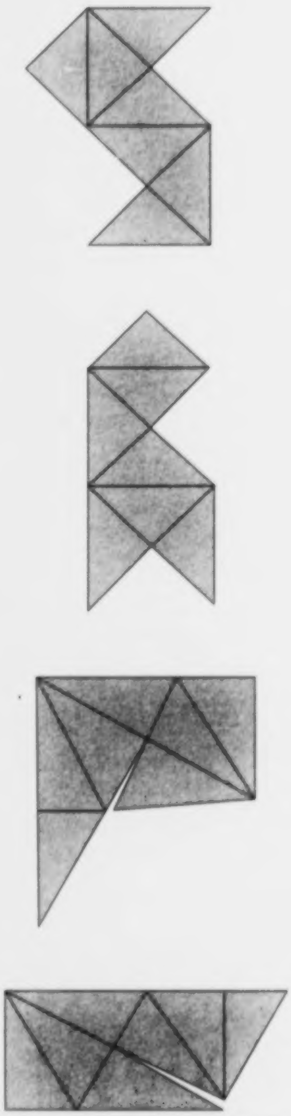


Figure 1. *The first two "drums" pictured above make the same "sound" although they are differently shaped. The same is true of the second pair.*

alike involves little beyond elementary linear algebra. "It's amazing how simple you can make this proof," muses DeTurck. It fits "on a postcard," he says. In fact, the mathematics department at Washington University did something like that: They had T-shirts made up with the proof on them.

But while the proof itself is simple, finding the pair of drums to begin with was not. It required insights culled from careful study of geometric analysis, as well as new theoretical techniques that involve a surprising range of mathematical disciplines, from the theory of partial differential equations to representations of finite groups. At one point it also had Gordon and Webb (who are married) filling up their living room with huge paper models of geometric drums. And in the final stages, it had them running up a sizeable phone bill with transatlantic calls and twice-a-day faxes.

First, though, what is a "mathematical drum," and why should there be any connection at all between the shape of a drum and the sound it makes?

The first part is easy enough to answer. A mathematical drum is just a shape in the plane—a region with an interior and a boundary—such as a circle, a square, an arbitrary polygon, or just a blob surrounded by a smooth curve. The "sounds" produced by such a drum are determined by the solutions of a partial differential equation known as the wave equation, which is used to describe any kind of wavelike phenomenon, from sound to light to water. In essence, the motion of a vibrating membrane (that is, a drum) is governed by this equation, together with the condition that the drum not vibrate on its boundary.

That condition is crucial. Physically, it just says that the drum is attached firmly to a frame. Mathematically, it restricts the set of solutions to the wave equation. Without some sort of boundary condition, a mathematical drum could make any sort of sound.

Among the solutions to the wave equation are certain ones that are purely periodic in time—that is, vibrations that produce a single, clear tone of a specific frequency. While it's the interior that does the actual vibrating, it's the boundary that determines which frequencies are allowed. These frequencies constitute the sounds a given drum can make. They depend solely on the drum's shape.

Kac's question asked whether or not that dependence could be turned around.

There were reasons to think it might. In 1911, Hermann Weyl proved one can hear the *area* of a drum. Weyl's result accords with the intuition that the bigger the drum, the lower the tone. Some years later, the Swedish mathematician Ake Pleijel proved one can hear the length of the boundary. And Kac himself conjectured—and I. M. Singer and Henry McKean proved—that the number of "holes" in a drum is audible. These results made it plausible that the sound of a drum might contain enough geometric information to specify the shape uniquely.

On the other hand, there were good reasons to think that wasn't the case. In particular, mathematicians started finding counterexamples in higher-dimensional generalizations of the problem. John Milnor, now at the State University of New York at Stony Brook, found the first counterexample in 1964, a pair of geometrically distinct, sixteen-dimensional "isospectral manifolds"—that being the fancy term for "sound-alike drums." Over the next two decades, other researchers found additional counterexamples in lower dimensions. But these discoveries seemed to have no systematic basis. It was as if they arose by accident.

That changed in 1985. Toshikazu Sunada of Nagoya University introduced a method that made it possible to construct examples of isospectral manifolds almost at will. Sunada's method gave rise to a veritable cottage industry of low-

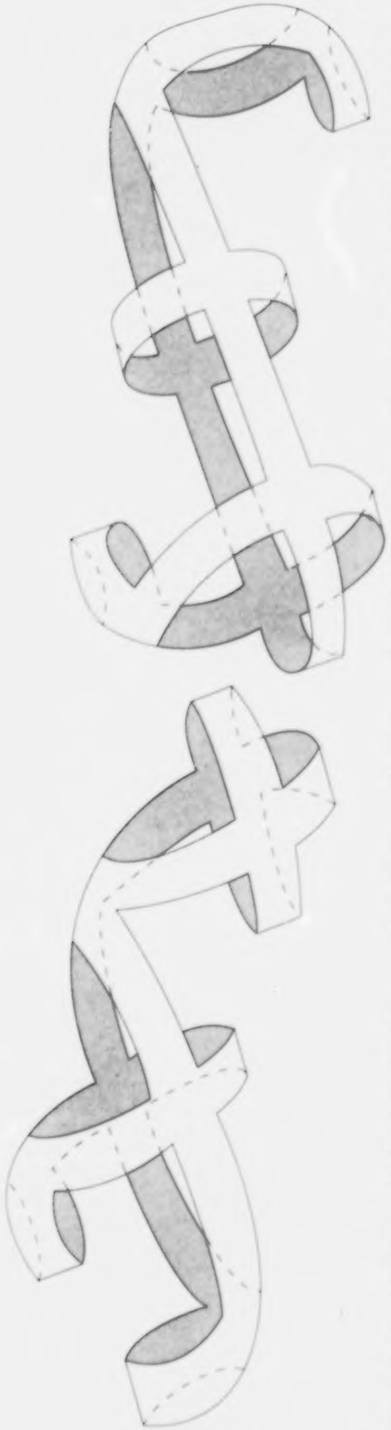


Figure 2. *A pair of Buser's isospectral "bells."*

dimensional examples, including surfaces that can actually be cut out of paper and assembled with tape. These surfaces fail to answer Kac's question only because they aren't flat but rather curve around in three dimensions, more like bells than drums. However, it was one of these bell-like pairs, an example cooked up by Peter Buser at the École Polytechnique Fédérale in Lausanne, Switzerland, that ultimately led to the long-sought solution of the original problem (see Figure 2 on page 14).

The inspiration came at a geometry conference at Duke University in March of 1991. Gordon showed a paper model of Buser's bell-like example in a survey talk. Wolpert was in the audience.

"Scott came up to me after the talk and said he'd noticed that these paper models had a symmetry to them, and if you 'modded out' by a symmetry—meaning simply smashing them down—then you got plane domains," Gordon recalls. "So he asked whether they were isospectral. And that's what led to all this. It really is just smashing them down."

Wolpert's hunch was right. But it took a while to find the proof. The shapes that result from flattening Buser's example are too complicated to compute their sounds exactly, so a direct comparison was impossible. Moreover, Sunada's method did not apply to the kind of surfaces, called orbifolds, that were required to make sense out of the flattening process.

However, help was already at hand. Pierre Berard at the University of Grenoble had generalized Sunada's method to one that worked in the orbifold setting. He had also introduced a crucial notion of "transplanting" solutions of the wave equation from one manifold to the other in an isospectral pair. Berard's results were exactly what the Americans needed.

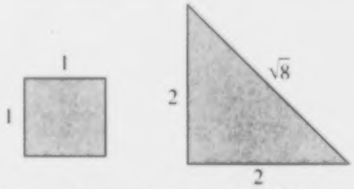
Even so, it took Gordon and Webb several weeks to find the right combination of ideas. They initially thought Buser's example was too simple to work, and spent days cutting out and taping together complicated paper models, looking for examples with good geometric properties. "We must have spent about a week, just building things out of paper," Gordon recalls. "And when we cut them in half, we must have spent about three hours trying to separate the two pieces [to get plane domains] before we realized they didn't separate!"

Finally they returned to the original example, which they had never completely abandoned. However, the last pieces of the proof came together while the two were thousands of miles apart—Webb at Dartmouth, Gordon in Germany. They hammered out the final details by phone and fax. By the time they got together, in Grenoble, they had a theory in place for a whole new class of orbifold-based isospectral manifolds, including a pair that lay flat in the plane, quietly answering No to Kac's old question.

That's not the end of the story, though. With help from Berard, Buser, and others, Gordon, Webb, and Wolpert identified the group-theoretic ideas that make the proof work, and now have a streamlined proof simple enough to fit on a T-shirt (see box on page 16). They also found other, simpler examples of sound-alike drums, some with as few as eight sides. Other researchers, including Peter Doyle and John Conway at Princeton University, have discovered additional shapes of elegant simplicity. (Using other methods, Conway and Neil Sloane at AT&T Bell Laboratories have found a family of four-dimensional examples similar to Milnor's original sixteen-dimensional example.)

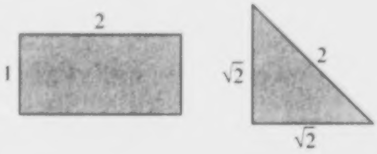
Gordon, Webb, and Wolpert's answer to Kac's question closes the book on one

Even if you can take infinitely many measurements with infinite precision, can you be sure of your conclusions? Or to put it differently, can two different objects look alike in every measurable way?



You Can't Hear the Shape of a Two-Piece Band.

Jon Chapman, a postdoc at Stanford University, took "scissors" to one of Gordon, Webb, and Wolpert's constructions and found a particularly simple pair of sound-alike drums, each consisting of two pieces. Chapman's two-piece drums, shown in the accompanying figures, are simple enough that it's possible to compute the exact sounds they make.



problem, but leaves others open — and raises new questions as well. "There are just tons of questions that come out of this answer," says an enthusiastic DeTurck. For instance, researchers right now only have examples of sound-alike pairs of drums; can there be sound-alike *triples*? Researchers also know that not all drums have isospectral twins (every circle, for example, makes its own, unique sound); is there a way to tell which drums do and which ones don't? And do the group-theoretic techniques of Sunada's method provide a unified explanation of all isospectral plane domains, or are there other ways of constructing sound-alike drums?

Finally, questions remain as to what implications the negative answer to Kac's question has for other inverse problems and their myriad applications. The one thing that seems clear is that the techniques mathematicians have developed in studying drums are likely to have repercussions throughout the rest of science.

Proof by Picture

Gordon, Webb, and Wolpert's first example of sound-alike drums came from a pair of curved surfaces designed by Peter Buser. Each drum consists of seven half-crosses glued together. Their streamlined proof that the drums are isospectral is based on group-theoretic principles plus Pierre Berard's "transplantation" technique, but the result is simple enough to be checked directly.

Figure 3 shows how to recombine pieces A–G of a standing wave on the seven half-crosses of the first drum into a standing wave on the second drum. For example, the combination $-B + C - E$ is formed by "flipping" pieces B and E upside down and adding them to piece C. The dark and dashed lines emphasize the required orientations.

This "transplantation" is easily seen to work both ways (it can be written as an invertible 7×7 matrix). All that remains is to check that the combinations fit together smoothly and are zero on the boundary. But this can be done piece by piece. For example, $-B + C - E$ is zero on the dark boundary because $-B$ cancels C there and E was zero to begin with, while it vanishes on the dashed boundary because B is zero there and C cancels $-E$; finally, $-B + C - E$ fits smoothly with $B + F - G$ on the diagonal because C-F and E-G already fit smoothly together while B is zero on the diagonal (and hence fits smoothly with its reflection).

The proof works just as well when the half-crosses are shrunk down to right isosceles triangles, and continues to work if the angles of the triangles are (simultaneously) changed. Thus one example gives rise to an entire family of sound-alike drums.

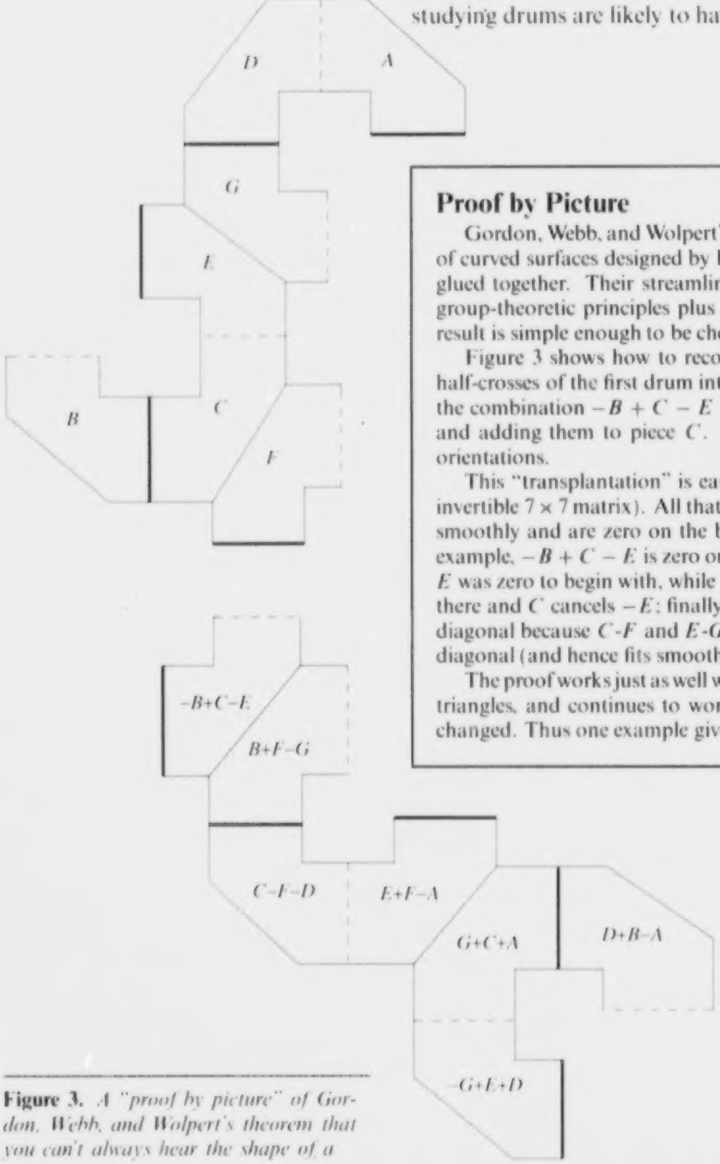


Figure 3. *A "proof by picture" of Gordon, Webb, and Wolpert's theorem that you can't always hear the shape of a drum.*

Environmentally Sound Mathematics

Among the crucial scientific issues of our age, few are as far-reaching as those posed by the environment. Researchers from all fields have been called upon to investigate and evaluate the effects human activities are having upon the earth, from the upper reaches of the atmosphere to the depths of the ocean. The complex web of relationships in the biosphere demands an interdisciplinary approach.

Rudimentary algebra and calculus are no longer enough to handle the sophisticated analyses that environmental scientists now know are necessary.

Long the preserve of biologists, chemists, oceanographers, meteorologists, and geologists, environmental science is now drawing more and more upon the expertise of mathematicians as well.

Researchers in environmental science have long made use of mathematics to one extent or another. What's new is the recognition that rudimentary algebra and calculus are no longer enough to handle the sophisticated analyses that environmental scientists now know are necessary. Advanced techniques in scientific computing and numerical analysis are coming to the fore as researchers tackle challenging problems ranging from acid rain to the effects of the world's oceans on global climate.

There is also growing interest in environmental science in the mathematics community itself. Recent meetings of the professional mathematics societies have featured presentations on environmental subjects, and last summer the Institute for Mathematics and its Applications, a mathematical think tank located at the University of Minnesota, held a four-week workshop on environmental modeling. Mathematicians are finding the field contains some interesting theoretical and computational problems. They are also finding a need in environmental studies for the ability of mathematics to build bridges between disciplines that are often separated by seas of jargon.

This kind of interdisciplinary work "takes time and energy, and it's not meant for everyone," says Mary Wheeler, a mathematician at Rice University, "but there are some really exciting challenges in it that will also drive some good results in mathematics."

Wheeler is one of the leaders in the movement of mathematicians into environmental science. She and colleagues at Rice and elsewhere have developed new mathematical tools for the study of fluid flow through porous materials. Their research combines the analysis of systems of nonlinear partial differential equations with sophisticated numerical algorithms that take advantage of new computer architectures such as massively parallel computation to solve the various equations. Among other applications, their efforts are aimed at helping environmental engineers plan remediation strategies for groundwater aquifers that have been contaminated by hazardous chemicals (see box on page 20).

One of the vexing aspects of environmental studies is the fact that the problems can span many scales of size. For example, a realistic climate model must consider everything from the microphysics of cloud nucleation to global circulation patterns. Likewise, plans for the isolation of nuclear waste must take into account physical processes occurring on a time scale of hours to months but keep an eye on safety standards valid for tens of thousands of years. In short, environmen-

Mary Wheeler. (Photo by Tommy Lavergne, Rice University.)

By working with mathematical models, environmental scientists can gain insights into systems that are too complex to study in any other way.

tal issues—take logging, for example—require researchers to look not just at the forest, but also at the trees (not to mention the spotted owls).

Mathematical modeling offers researchers the opportunity to identify and clarify mechanisms that connect phenomena at different scales, Wheeler says. In some cases sheer computational power makes the connection possible; in other cases, the mathematical equations themselves reveal the crucial interactions. By working with mathematical models, environmental scientists can gain insights into systems that are too complex to study in any other way.

Computers, and the high-tech algorithms that run on them, are making it possible to do the calculations required by these sophisticated mathematical models. "More and more people are recognizing that, with these tools, we can solve very complex problems," says Julius Chang, an atmospheric scientist at the State University of New York at Albany. Researchers no longer have to rely on unrealistic simplifications in order to make the computations tractable. "We can tackle many problems head on," Chang says.

Chang's group, for example, has developed an acid-rain model called RADM (for Regional Acid Deposition Model) which includes a system of coupled differential equations for a set of sixty different chemical species (see Figure 1). These aren't your nice, neat, textbook equations, either. Printed out, a typical RADM equation runs on for line after line of cryptic symbols and mixed upper- and lower-case letters, and could easily be mistaken for an old-fashioned computer core dump. RADM's equations take into account effects such as atmospheric advection and mixing, gas-phase chemical reactions, cloud mixing and "wet scavenging," dry deposition (acid can "fall" even when it isn't raining), and the location of sources of various pollutants (what comes down must have gone up).

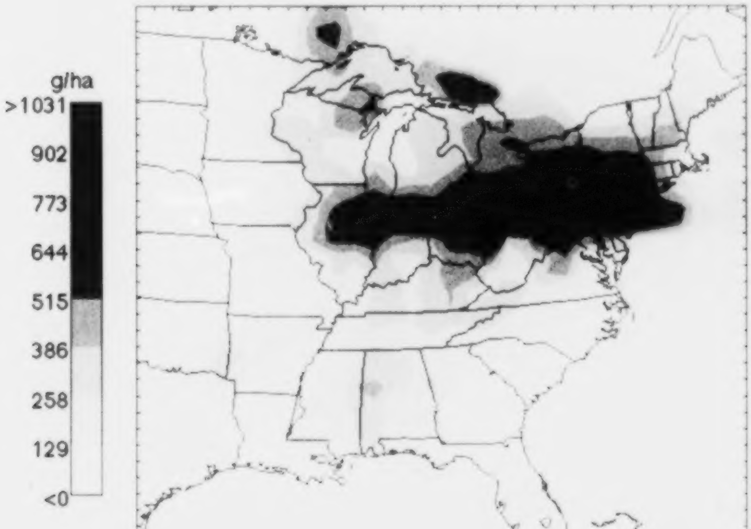


Figure 1. *Cumulative wet sulfur deposition (in grams/hectare) for eastern North America over a three-day period in April 1981 as calculated by the Regional Acid Deposition Model (RADM) using actual reported emission rates of various atmospheric pollutants. (Figure courtesy of Julius Chang.)*

Even so, there are limits to what computers can do. Take clouds, for example. "A cloud serves as an elevator for pollutants," Chang explains. But clouds are notoriously difficult to model. Scientists specializing in cloud processes have created "wonderfully detailed" models of cloud formation, Chang says, but their models "are too computationally complex to use on a regional scale. A single cloud can fill up a whole computer." So for now, RADM makes do with a cloud model that's fairly realistic but highly simplified.

If clouds are hard for computers to handle, oceans are even worse. There's a huge gulf between what goes on on the open seas and what even the largest supercomputer can model. That's a problem because oceans play a significant role in determining climate—and climate is one of environmental science's biggest concerns.

"If you're interested in climate, you're interested in the ocean," says Mac Hyman, a mathematician in the Theoretical Division at the Los Alamos National Laboratory in New Mexico.

The problem is not that oceans are hard to understand; in some respects, their equations are simpler than those of the atmosphere. Basically, the earth's oceans act as a gigantic heat reservoir and transport system. They exert a long-term influence on global climate by absorbing and emitting heat and carbon dioxide. The equations that describe all this are pretty well worked out. They include the Navier-Stokes equation, which underlies all fluid flow problems, and other partial differential equations describing gas exchange and heat transfer. All told, the basic equations of ocean dynamics can be written down on a single page. The problem is, these equations can't be accurately solved on existing computers—at least not by standard numerical methods.

That's because most of the kinetic energy of the ocean is found at scales too small for standard models to resolve, explains Hyman's colleague Darryl Holm. Unlike the atmosphere, where length scales on the order of 100 kilometers dominate the dynamics (just think of storm fronts), much of the ocean's energy exists in the form of eddies and waves that are five to ten times smaller.

"Without modeling some aspects of the small-scale, high-frequency waves and eddies, we can't know whether our global oceanic models are truly reliable," Holm says.

But researchers can't just tell the computer to take a closer look at the small-scale phenomena—or rather, says Hyman, "You can't afford the computer costs to resolve them. If you go down a factor of 10, that's a factor of 10 in three dimensions, so that's a factor of 1000—plus a factor of 10 in time, so that's a factor of 10^4, in computer costs, and no one's talking about those kinds of gains [in computer technology] in the next few years."

However, Holm and Roberto Camassa, also at Los Alamos, have developed some new mathematical approaches that may get around the problem. Their basic idea is to simplify the equations for ocean dynamics by taking advantage of the fact that certain important parameters, such as the ratio of surface wave amplitude to ocean depth and the ratio of depth to width of the ocean, are extremely small. If done carefully, the simplified equations will reliably represent the average effects of the high-frequency, small-scale elements on the large-scale dynamics. What should come out, says Hyman, is the correct average answer, "which is what we're looking for in the climate anyway."

Hyman, Wheeler, and others see a permanent role for mathematicians in en-

Mathematical modeling offers researchers the opportunity to identify and clarify mechanisms that connect phenomena at different scales, Wheeler says.

vironmental science. After all, one of the things mathematicians do is to solve problems. And when it comes to the environment, the problems seem to be getting bigger all the time.

Mathematical modeling not only may help plan the cleanup of contaminants, Wheeler adds, it may also help contain costs.

Bugs in the Program

The cleanup of underground aquifers contaminated by hazardous chemicals is serious—and costly—work. For example, the disposal of carbon tetrachloride over a period of eighteen years at the Hanford Site in south central Washington state has left contaminated groundwater over a five-square-kilometer area; cleaning it up could cost as much as $300 billion. And that's just one particularly bad example. The problem is not limited to a few locations. In 1986, the Environmental Protection Agency estimated there to be leaks in as many as 35% of the roughly 800,000 gasoline storage tanks in the U.S., with more than half reaching the water table.

"Contamination of aquifers by polluted streams and ponds, leaking storage tanks, agricultural chemicals, gasoline spills, and dumping has become a serious and widespread threat to public health," says Mary Wheeler. Wheeler has taken a keen interest in developing mathematical models that can assist environmental engineers plan their cleanup strategies.

One such strategy is known as *in situ* biorestoration. The basic idea is very simple: Certain microorganisms will actually digest or otherwise remove contaminants such as carbon tetrachloride—but only if encouraged to do so by the introduction of dissolved oxygen or other triggering nutrients. What makes it complicated are the complex interactions of groundwater, contaminant, organisms, and nutrients, which are flowing through material that may itself be highly heterogeneous.

That's where Wheeler's mathematics comes in. Wheeler and colleagues have developed mathematical models that describe these interactions in terms of nonlinear partial differential equations. They have also developed new computational techniques to solve these equations numerically and display the results using three-dimensional computer graphics. While noting there's still a lot of work to be done, Wheeler says these models should give researchers some much-needed insight into what's going on at contamination sites and how biorestoration can be used to best effect.

Mathematical modeling may not only help plan the cleanup of contaminants, Wheeler adds, it may also help contain costs. That's because "experiments" run on a computer are much cheaper than actual field experiments—and some experiments can only safely be tried in a computer "environment." The cost of conducting field experiments can run into the millions of dollars, Wheeler notes. "Doing it on the computer is very cheap."

Fighting over who gets to write down the next term in the equation: Darryl Holm (left), Roberto Camassa (center), and Mac Hyman. (Photo by Fred Rick, Los Alamos National Laboratory.)

Disproving the Obvious in Higher Dimensions

Not everything that's "obvious" is necessarily true.

Scientists in all disciplines know that drawing "obvious" conclusions, even from well-founded facts, is a dangerous game, unless those conclusions can be backed up by experimental verification. The same is true in mathematics, except that mathematical proof takes the place of laboratory experimentation. Mathematical explorations are guided by intuition, but only when their intuitions are confirmed by proof do mathematicians accept the "obvious" as true. This approach is necessary because sometimes what seems "obvious" just ain't so.

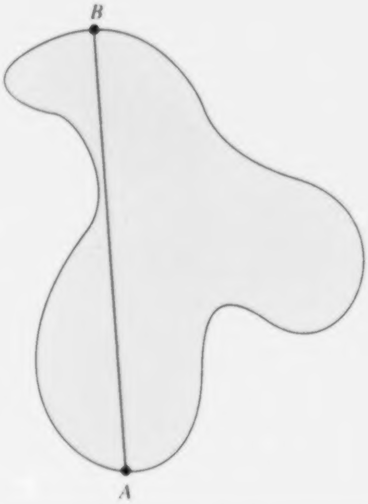


Figure 1. *A and B are the farthest apart of any two points in the shaded region. The distance between them is called the "diameter" of the region.*

Mathematicians saw that happen not once, but twice in 1992. In similar but separate developments, researchers discovered that two facts from plane and solid geometry, facts that cry out for obvious generalization to geometric figures in *any* dimension, do not hold in that kind of generality. Their findings reaffirm researchers' suspicion that ordinary spatial intuition is not up to the task of thinking in higher dimensions.

That would be of only academic interest were it not for the fact that higher-dimensional geometry plays an important role in many mathematical applications. "I've been asked questions about higher-dimensional geometry by people who are interested in speech recognition and by people who are interested in algorithms for dealing with DNA," says Peter Shor, a research mathematician at AT&T Bell Laboratories in Murray Hill, New Jersey. Higher-dimensional geometry provides a natural mathematical framework for dealing with problems involving several variables or long strings of data. In particular, it has figured prominently in the development of so-called error-correcting codes, which are mathematical constructions that underlie the reliable storage and transmission of data in satellite telemetry, computer modems, and even compact disks.

One of the "obvious" generalizations was a problem that had been bothering mathematicians for the better part of sixty years before it was tackled by Jeff Kahn at Rutgers University in New Brunswick, New Jersey, and Gil Kalai at the Hebrew University in Israel. In 1933, the Polish mathematician Karl Borsuk proved that any region in the plane whose "diameter"—the largest distance between two points in the figure—is equal to 1 can be cut into three pieces, each of diameter strictly less than 1 (see Figures 1 and 2). This generalizes the completely trivial observation that a one-dimensional figure of diameter 1—that is, a line segment of length 1—can be cut into two shorter pieces.

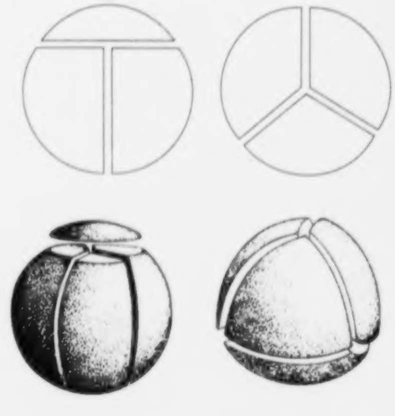


Figure 2. *Two ways to partition the circle and sphere into pieces of smaller diameter.*

On the basis of these two cases, Borsuk asked the obvious question: Is it always possible to cut any d-dimensional shape of diameter 1 into $d+1$ pieces each of diameter less than 1? The "obvious" affirmative answer came to be called Borsuk's conjecture.

For many figures, of course, the task takes fewer than $d+1$ pieces. The square of diagonal 1, for example, can be cut neatly in half. On the other hand, an equilateral triangle in the plane, a tetrahedron in space, and their cousins in higher dimensions definitely do require $d+1$ pieces: Since their vertices are all mutually a unit distance apart, each vertex must go into a separate piece. Borsuk's conjecture

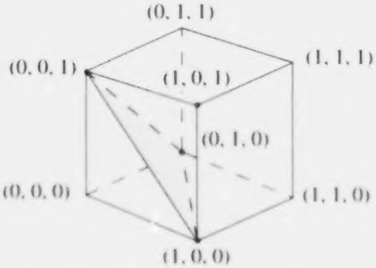


Figure 3. *A diagonal "slice" through the unit cube in three dimensions.*

One Plus One Equals 1.1?

If Borsuk's conjecture were true, then it should apply to any geometric figure. In particular it should be true for the vertices of a d-dimensional "slice" through the unit "cube" in $(d + 1)$-dimensional space. The coordinates of the vertices of the unit cube are all zeros and ones, and one way to slice through it is to restrict to vertices that have a specified number of ones (see Figure 3).

Each vertex can be thought of as specifying a subset of the integers $\{1, 2, \ldots, d + 1\}$ according to which coordinates are ones and which are zeros. For example, $(1, 1, 0, 0, 1)$ specifies the subset $\{1, 2, 5\}$. Under this interpretation, the distance between two vertices is related to the size of the intersection of the corresponding subsets: The smaller the intersection, the greater the distance.

In this setting, David Larman observed, Borsuk's conjecture reduces to a combinatorial assertion about sets: If S is a family of subsets of $\{1, 2, \ldots, d + 1\}$ such that all the sets in S have the same number of elements and such that every two sets in S have at least n elements in common, then S can be partitioned into $d + 1$ parts so that in each part every two sets have at least $n + 1$ elements in common.

It was this version of Borsuk's conjecture that Kahn and Kalai found to be false. The vehicle they used to get there is a theorem of Frankl and Wilson: Let k be a power of a prime number, and let S be a family of subsets of $\{1, 2, \ldots, 4k\}$, each with $2k$ elements, such that no two members of S have k elements in common. Then S has at most $2\binom{4k-1}{k-1}$ members. (This bound is less, by an exponential factor, than the total number of sets with $2k$ elements.)

The reader is invited to ponder just how the Frankl-Wilson theorem contradicts Borsuk's conjecture. But remember Kalai's warning: "It's an example of an extremely short proof that was quite difficult to find."

doesn't say that *every* figure needs to be cut into $d + 1$ pieces in order make all the pieces have smaller diameter; it just says $d + 1$ is the *most* you ever need.

Things started looking good for Borsuk's conjecture in 1946, when the Swiss mathematician Hugo Hadwiger showed that any d-dimensional geometric figure can be cut into $d + 1$ pieces of smaller diameter, if its boundary is smooth. In other words, Borsuk's conjecture is true for things like the d-dimensional sphere—shapes without corners or creases.

Then in 1955, the English mathematician H.G. Eggleston proved Borsuk's conjecture for $d = 3$. And that's pretty much where things stood until 1992, when Kahn and Kalai came in and knocked Borsuk's conjecture flat on its back.

Actually, Kahn and Kalai have not ruled out Borsuk's conjecture altogether; it might still be true in quite a few more dimensions. What they showed is that, for high dimensions, the minimum number of pieces required to cut any d-dimensional object into pieces of smaller diameter grows much more rapidly than $d + 1$. Specifically, Kahn and Kalai proved that the minimum exceeds $1.1^{\sqrt{d}}$.

That formula doesn't do much good for small values of d, where $d + 1$ is larger than $1.1^{\sqrt{d}}$. But starting around $d = 10{,}000$, the Kahn-Kalai bound kicks in. (The first instance where $1.1^{\sqrt{d}}$ is greater than $d + 1$ occurs at $d = 9162$. With a little more care, Kahn says, they can obtain a formula that works down around 2000.) More to the point, their result shows that Borsuk's conjecture is *badly* wrong at high dimensions. The number of pieces needed grows exponentially, not linearly.

The form of the result might seem to suggest a long, complicated proof. After all, square roots don't often appear as exponents, and 1.1 is not the most natural number in the world. In fact, the proof is surprisingly short. It's only a few lines long. That doesn't mean the proof was easy to come by, though. "It's an example of an extremely short proof that was quite difficult to find," says Kalai.

Jeff Kahn. (Photo by Nick Romanenko, Rutgers University.)

Peter Shor and Jeff Lagarias. (Photo courtesy of AT&T Bell Labs.)

The proof is based on two ideas. The first, due to David Larman at University College in London, is an interpretation of Borsuk's conjecture as a statement about families of finite sets and their intersections. The second is a theorem due to Peter Frankl at the Centre National de la Recherche Scientifique in Paris and Richard Wilson at the California Institute of Technology about the size of such families (see box on preceding page). The hard part was "figuring out what to do with these ideas," Kahn recalls. Once they found the right construction, though, the contradiction to Borsuk's conjecture was an immediate consequence of the Frankl-Wilson theorem.

While it wipes out Borsuk's conjecture in general, Kahn and Kalai's construction of counterexamples leaves a lot of dimensions unaccounted for. In particular, "for dimension four, you clearly need a different way to look at the entire problem," says Kalai. The conjecture could be true or it could be false in that case. Nobody knows. And it could be another sixty years before anyone finds out. Or another six hundred years. Or it could be proved tomorrow.

Kahn and Kalai's cutting apart of Borsuk's conjecture was actually the second of the two counterintuitive geometric discoveries of 1992. Earlier in the year, Peter Shor and Jeff Lagarias, also at Bell Labs, took on another sixty-year-old problem, one with roots even older than that. The conjecture they looked at is based on a simple observation about squares in the plane: If you try to tile the plane with squares of equal size, then you necessarily wind up with squares that have an entire side in common. In fact there's essentially only one kind of tiling of the plane by squares, namely a checkerboard tiling in which the rows have been shifted by arbitrary amounts (see Figures 4a and 4b).

If you get your hands on a set of child's building blocks, you can convince yourself that something similar is true in three dimensions: If you "tile" space with cubes of equal size, you wind up with cubes that have an entire side in common. In this case, of course, the common side is a two-dimensional square.

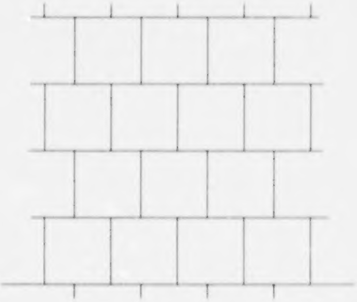


Figure 4a. *A tiling of the plane by unit squares.*

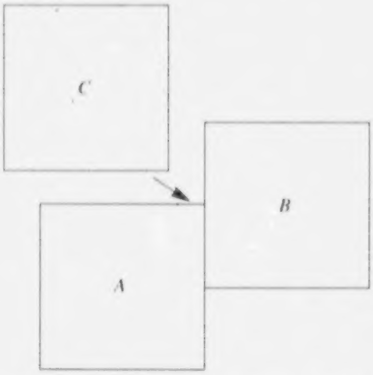


Figure 4b. *When square C is moved into the corner, it will have an entire side common with square A.*

BEST COPY AVAILABLE

What is clear, from both recent results, is that geometric intuition is a deceptive guide.

“Tiling space,” whatever the dimension, means filling the entire space without any overlapping.

In 1930, the German mathematician Ott-Heinrich Keller took the plunge. He conjectured that no matter what the dimension d, if you tile d-dimensional space with d-dimensional “cubes” of equal size, then you wind up with cubes that have an entire $(d-1)$-dimensional “side” in common.

Actually, Keller was just generalizing a conjecture of Hermann Minkowski, who in 1907 made the same observation, but restricted to “lattice tilings”—tilings for which the centers of the cubes form a regular, grid-like lattice of points, like the locations of carbon atoms in a crystal of diamond. As it turns out, Minkowski was right and Keller was wrong.

Oskar Perron made his countrymen look good in 1940, proving Keller's conjecture in dimensions up to six. Two years later, the Hungarian mathematician György Hajós vindicated Minkowski completely, showing the original conjecture for lattice tilings is true in all dimensions. That left Keller's more general conjecture in dimensions seven and up. The issue lay unresolved for fifty years.

No more. Lagarias and Shor have found an explicit counterexample to Keller's conjecture in ten-dimensional space. This also kills the conjecture in dimensions eleven and up, because as soon as Keller's conjecture fails to be true at one dimension, it automatically stops being true at all higher dimensions. (A tiling by d-dimensional cubes can be converted into a layer of $(d+1)$-dimensional cubes, and then copies of the layer can be stacked to fill all of $(d+1)$-dimensional space, with the layers shifted so that there are no entire sides in common between layers.) The only unresolved cases are dimensions seven, eight, and nine.

In a sense, those cases require only patience—and maybe a high-speed computer the size of a major galaxy. That's because in addition to proving Keller's conjecture for dimensions up to six, Perron also showed how the conjecture could be checked in any given dimension by looking at a finite number of different tilings: If no counterexample is found in this finite set, then the conjecture is true (in that dimension). Unfortunately the number of tilings to be checked is unbelievably large: 2^{2^d}. No one in his right mind, and no mathematician either, would set out to sort through 2^{128} possible tilings to check the case $d = 7$.

Nevertheless, Lagarias and Shor did something of the sort to find their ten-dimensional counterexample. They based their construction on work of Kereszyély Corrádi at Eötvös Loránd University and Sándor Szabó at the Technical University in Budapest, who two years earlier had introduced a new approach to looking for counterexamples. By studying the output of limited computer searches, Lagarias and Shor found tilings in dimensions three, four, and five that almost gave counterexamples in those dimensions. By cobbling these near-misses together, they manufactured legitimate counterexamples, first in dimension twelve, and then in dimension ten.

It's unclear if the same techniques can be brought to bear in dimensions seven, eight, and nine. Lagarias and Shor say it's possible the conjecture may fail even in dimension seven, but the counterexamples are too structureless to find. “The amazing thing is that there actually existed a counterexample that had a simple enough structure that you could actually find it,” says Lagarias.

It's also unclear if their counterexample to Keller's conjecture will have any direct applications to things like error-correcting codes—so don't expect next year's line of CD players to be based on a tiling of ten-dimensional space. However, Lagarias notes, the cube-tiling constructions give rise to novel types of "nonlinear" codes quite unlike the linear codes that are used in current applications.

What is clear, from both recent results, is that geometric intuition is a deceptive guide. "High-dimensional space is very strange," says Lagarias. Adds Shor: "If you're going to make conjectures about high dimensions, you should use some basis other than just extrapolation."

In fact, Shor goes so far as to make his own conjecture about higher-dimensional geometry: "Conjectures based solely on low-dimensional examples are false in high dimensions," he asserts. Asked if that includes his own conjecture, Shor amends the statement: "Conjectures based solely on low-dimensional examples are *likely* to be false."

Here's Looking at Euclid

The latest results involve some pretty highfalutin math, but not all counterintuitive results in higher-dimensional geometry are hard to prove. Here's one you can "see" for yourself.

Start by drawing four circles of radius 1 centered at the points $(1, 1)$, $(1, -1)$, $(-1, 1)$, and $(-1, -1)$ and then add a fifth circle centered at the origin and touching the other four (see Figure 5). This central circle is clearly contained in the square around the four outer circles.

The same thing is true in three dimensions: If eight spheres of radius 1 are centered at the points $(\pm 1, \pm 1, \pm 1)$, then a ninth, central sphere touching them all stays within the cube around the eight (see Figure 6).

It would seem obvious that no matter what the dimension, the central "sphere" always stays within the corresponding d-dimensional "cube." It's just not true.

Here's why. By the (generalized) Pythagorean theorem, the distance from the origin to any of the centers of the outer spheres is

$$\sqrt{(\pm 1)^2 + (\pm 1)^2 + \cdots + (\pm 1)^2} = \sqrt{d},$$

and consequently the radius of the central sphere is $\sqrt{d} - 1$. But the distance from the origin to any side of the cube is always just 2. So when $d = 9$, the central sphere touches each side of the cube, and for $d \geq 10$ it pokes outside the cube.

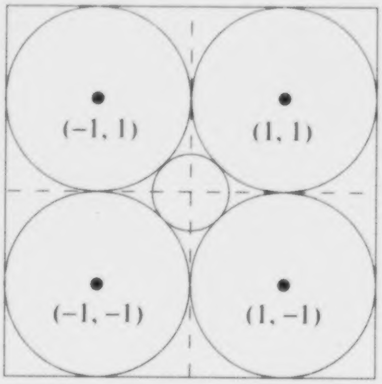


Figure 5. *The small inner circle touches all four circles of radius 1 and stays within the square.*

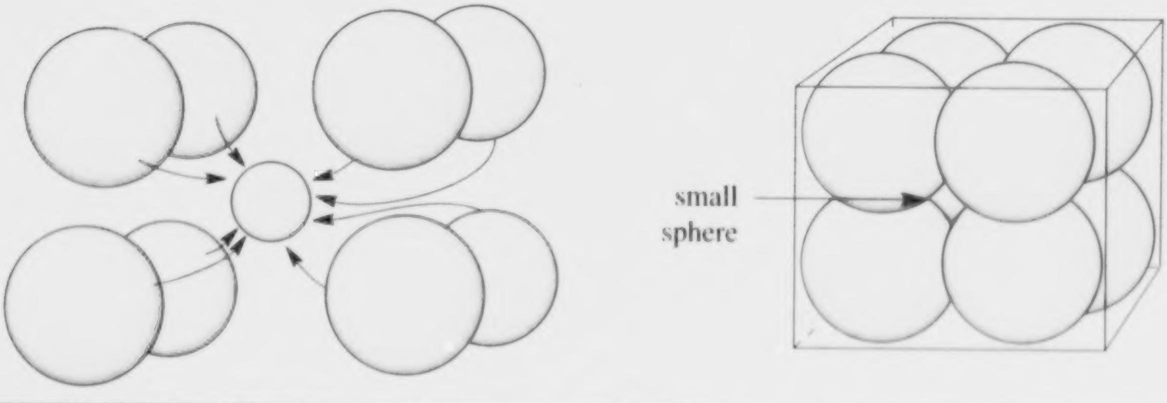


Figure 6. *Similarly, in three dimensions, the small inner sphere, which touches all the larger ones, remains inside the cube. This is no longer true in higher dimensions!*

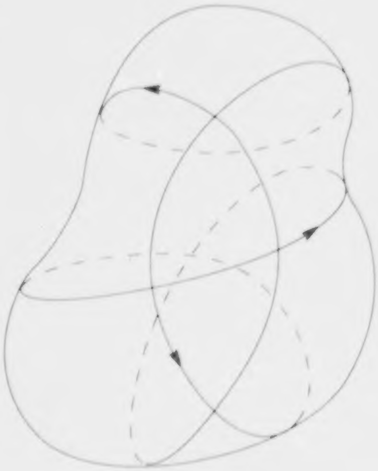


Figure 1. *A closed geodesic on a distorted sphere.*

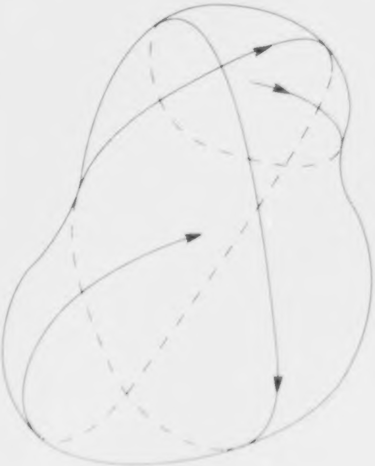


Figure 2. *A nonclosed geodesic on a distorted sphere.*

29 BEST COPY AVAILABLE

Collaboration Closes in on Closed Geodesics

Individually, neither hydrogen nor oxygen can combust. But put together and ignited by a spark, they are capable of exploding with enough power to propel a rocket into outer space. In mathematics, something similar can happen when two theories are brought in contact and set off by the spark of a new idea. Recently two mathematicians with expertise in separate specialties joined forces to solve a problem in differential geometry that had been on the books for more than sixty years.

Victor Bangert, at the University of Freiburg in Germany, and John Franks, at Northwestern University, have shown that no matter how badly you distort a sphere, there will always be infinitely many "closed geodesics" on it: rubber band-like curves that are determined by the curvature of the distorted surface (see Figure 1). The previous best result had been that every distorted sphere had at least *three* such geodesics—and that theorem dates back to the 1920s.

The new result is mainly of theoretical interest, but that doesn't mean it won't ever find practical applications. According to Robert Molzon, program director for geometric analysis in the Division of Mathematical Sciences at the National Science Foundation, differential geometry is applicable to "everything from general relativity and understanding the large-scale structure of the universe down to very small-scale problems such as boundaries between phases [e.g., liquid and gas] in materials science." Bangert and Franks's theorem is one more tool with which to study such problems.

Molzon is also encouraged by the new collaboration between two seemingly disparate mathematical areas: differential geometry and dynamical systems. "Bringing together these two areas is a big step," he says. Bangert and Franks solved the geodesic problem through a "divide and conquer" approach, with Bangert using classical techniques in differential geometry on one part of the problem and Franks bringing dynamical systems theory to bear on the other part.

Whether it's concerned with applications to relativity theory or materials science, or with more abstract issues in mathematics itself, differential geometry can be loosely described as the study of curvature. Geodesics are among its fundamental objects. A geodesic is basically just a path that follows the curvature of whatever surface or space it lies in. The precise definition implies that geodesics have a "shortest path" property. In particular, the shortest path between two points always lies along a geodesic.

The geodesics on a perfect (i.e., undistorted) sphere are the great circles, such as the equator or any line of longitude on the globe. Every one of them is closed. But as soon as you hammer on the sphere, that's no longer true. In general, when a geodesic traveling in one direction approaches a bump or a dent, it gets deflected in some other direction, much as a golf ball may veer away from the cup on an uneven green. It can easily happen that the geodesic will never find its way back to where it started (see Figure 2).

When there are bumps and dents everywhere, it's possible to imagine *every* geodesic wandering about endlessly. But that doesn't happen. George David Birkhoff proved in 1917 that every distorted sphere has at least one closed geodesic.

John Franks.

Whether it's concerned with relativity theory, materials science, or more abstract issues in mathematics itself, differential geometry can be loosely described as the study of curvature.

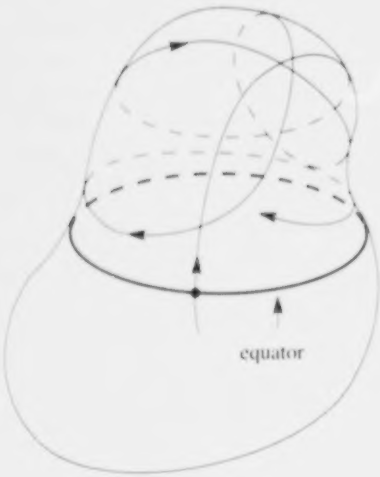


Figure 3. *A geodesic that gets "trapped" in the "Northern hemisphere."*

Twelve years later, two Russian mathematicians, Lazar A. Lyusternik and Lev Schnirelmann, went a step further. They proved there are always at least *three* closed geodesics on any distorted sphere.

More than half a century went by without a single closed geodesic being added to the count. Then Bangert had an idea for renewing Birkhoff's original attack on the problem. Part of the problem, Bangert saw, could be dealt with using techniques coming purely from differential geometry, his own specialty; the rest would require results from the theory of dynamical systems—and for that he sought the expert help of Franks.

"Bangert really kindled my interest in clarifying exactly what was needed to get this result," Franks recalls.

At first glance, differential geometry seems a far cry from the theory of dynamical systems. One subject is concerned mainly with objects that are fixed and permanent, while the other, virtually by definition, is interested in how things change. But the two aren't complete strangers. Birkhoff had already shown how the geodesic question could be translated into a problem purely in dynamical systems.

Birkhoff's translation starts with a single closed geodesic that loops once around the sphere without intersecting itself. This curve acts as a kind of "equator" separating two "hemispheres." Any other geodesic that crosses the equator either continues to cross it infinitely often (which is the case for closed geodesics if you keep following them around and around), or else it eventually gets trapped in one hemisphere (see Figure 3).

The former case, Birkhoff showed, leads to a dynamical system. Each crossing of the equator can be described by two parameters: one for the location of the crossing (i.e., its "longitude"), and one for the angle it makes with the equator. These parameters can be plotted on a washer-shaped region known as an annulus (see Figure 4). So each crossing of the equator by a geodesic corresponds to a point in the annulus, and, conversely, each point in the annulus corresponds to a crossing of the equator by some geodesic.

The theory of dynamical systems enters in when you follow geodesics from one crossing to the next. This defines a map of the annulus back onto itself—and maps from a region back to itself are one important kind of dynamical system. In particular, such maps can be iterated (that is, applied repeatedly). In the case of Birkhoff's annulus map, this corresponds to following a geodesic from one crossing to the next. The key point is that periodic points for Birkhoff's annulus map—that is, points on the annulus that eventually get mapped back onto themselves—correspond to closed geodesics. So to show there are infinitely many closed geodesics on a distorted sphere, it's enough to prove that Birkhoff's annulus map has infinitely many periodic points.

Bangert saw a division of labor. First of all, something had to be done in the case when Birkhoff's annulus map is not defined, which can happen, for instance, when a geodesic crosses the equator and gets trapped in the northern hemisphere. Bangert handled this case using classical techniques in differential geometry. In fact, his proof implies there are infinitely many closed geodesics anytime there are two geodesics that don't cross each other at all.

It remained to prove that, when the annulus map is defined, it's guaranteed to have infinitely many periodic points (corresponding to infinitely many closed geodesics). This case Bangert left to Franks, an expert on annulus maps.

At first glance, differential geometry seems a far cry from the theory of dynamical systems. One subject is concerned mainly with objects that are fixed and permanent, while the other is interested in how things change.

Birkhoff's annulus map, it turns out, has a special property: When it maps the annulus back onto itself, it preserves the area, if not the shape, of any piece of the annulus. Birkhoff himself used this feature to prove that, at least under certain circumstances, his annulus map would have a fixed point, corresponding to a closed geodesic that intersects the "equator" in only one point. The proof, however, had nothing to do with geodesics or differential geometry; it was pure dynamical systems, a statement about area-preserving annulus maps.

Area-preserving annulus maps have been a staple of dynamical systems theory ever since. Franks had proved a generalization of Birkhoff's theorem (more properly called the Poincaré-Birkhoff theorem), and this was why Bangert approached Franks, in 1988, for help on the geodesic problem. It was clear what needed to be proved, Franks recalls. It just wasn't clear—at first—how to prove it.

Finally, in 1991, it became clear. Franks's theorem says that *any* area-preserving annulus map either has no periodic points, or else it has infinitely many of them. For the geodesic problem, the no-periodic-point possibility can be ruled out, and that leaves the long-sought conclusion: The sphere, no matter how badly distorted, still has an infinite family of closed geodesics.

Franks's theorem and Bangert's analysis don't completely close the book on the closed geodesic problem. If anything, the fact that there are always infinitely many closed geodesics raises a host of new questions. There are also questions raised by the proof itself. For example, among the closed geodesics on a distorted sphere, is there always one for which the Birkhoff map is defined? (If that's the case, then Franks's theorem alone would complete the proof that there are infinitely many closed geodesics.) The list of potential problems and new questions runs on—as endlessly as the geodesics themselves.

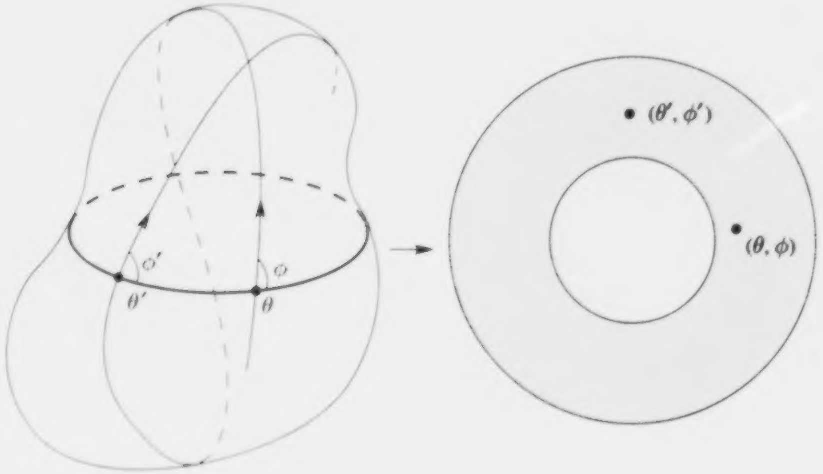


Figure 4. *The geodesic crosses the "equator" at a certain point along the equator (indicated by θ) and at a certain angle (indicated by ϕ). The values of θ and ϕ at any crossing are plotted onto the annulus.*

Rob Almgren and Andy Roosen at the Geometry Center in Minneapolis. (Photo by Barry Cipra.)

Crystal Clear Computations

Rob Almgren and Andy Roosen have a spirited competition going on. They're trying to see who can grow the nicest-looking crystals. Almgren and Roosen aren't working with chemicals in a laboratory, though. Instead, the crystals they grow are geometric shapes that develop in the purely numerical environment of a computer.

The two mathematicians—Almgren at the University of Chicago, Roosen a graduate student at Rutgers University—are members of a computational crystal growing group. The informal network of researchers, headed by Jean Taylor at Rutgers and Fred Almgren (Rob Almgren's father) at Princeton University, is part of a new trend in mathematics to combine the power of mathematical analysis with the speed and versatility of modern computers to tackle complex problems of fundamental importance head on.

Jean Taylor. (Photo by Rebecca Savoie.)

The computational crystal growers are creating mathematical models and analytic techniques that will give scientists powerful new tools for studying the nature of crystals. These new tools, researchers say, will help accelerate the future design of materials with special properties of strength, "shape memory," and even superconductivity. At the same time, the computational crystal studies raise challenging problems in pure mathematics and numerical analysis, the solutions to which may well have applications in other, unrelated areas. The subject contains a "wealth of new geometric phenomena," says Taylor. "There are all these things out there waiting to be explained."

Snowflakes, for one. Just how the familiar six-sided crystal takes shape is still largely a mystery. Scientists know that the final shape depends on the conditions of temperature and supersaturation of water vapor while the snowflake is forming, but "the exact mechanisms are far from clear," according to Rob Almgren. The computational crystal growers will know they're on to something when they are able to mimic the growth of snowflakes.

But other applications are likely to come first. While snowflakes have intrinsic scientific (and aesthetic) appeal, there are also practical considerations driving research in crystal growth. The strength of steel, for example, is determined in part by the way crystals form as the metal cools from an initial, liquid state. Likewise, many properties of semiconductors depend on the way impurities in the melt are "driven out" in the process of solidification. Crystal growth is also crucial to the up-and-coming manufacturing technique known as molecular-beam epitaxy, in which materials are created one atomic layer at a time in a kind of ultra-high-tech version of spray painting.

Those phenomena all involve the growth of dendrites, and that's what Almgren and Roosen have been vying to recreate on their computer screens. Dendrites are structures that branch in complicated ways (the name comes from the Greek word for "tree"). In crystals, they are created by the interplay between surface energy and diffusion of heat or chemical impurities.

Surface energy in crystals is closely related to the area-minimizing surface tension that tends to keep soap bubbles and raindrops spherical. However, for crystalline materials the energy of a piece of surface depends on the direction it faces. Minimization of this "anisotropic" surface energy pulls the crystals into nonspherical shapes such as the cubical crystals of table salt. Heat and excess

The computational crystal growers are creating mathematical models and analytical techniques that will give scientists powerful new tools for studying the nature of crystals.

The theoretical results "might never have been obtained without the new ideas generated by trying to do numerical computation," Rob Almgren notes.

chemical concentrations come from the release of latent heat and impurities whenever a portion of the liquid crystallizes.

"When heat is released, it must diffuse away," Roosen explains. Until it does, the region where it was released is too warm to crystallize further. "But what this means is that little bumps sticking out are able to diffuse their heat away faster than little dips, consequently they grow faster," he adds. "Without surface energy, this would happen with any bump or dip no matter how small, so the crystal would develop arbitrarily small structures. With surface energy, there is a lower limit on how small the fingers poking out into the cold can be." In other words, "release of latent heat and diffusion creates instability, surface energy controls it," Roosen says. The net result is the rapid advance of stable dendritic "tips" and the creation of characteristic branching patterns.

While that description seems straightforward enough, formulating it mathematically and then turning the equations into workable computer algorithms is a different matter entirely. Even in a simplified, two-dimensional setting, no one has yet come up with a method to match the rich range of structures seen in real experiments. Says Taylor: "Various people have announced that they 'understand' dendrites. I don't."

Almgren and Roosen's two-dimensional pictures look promising, however (see Figures 1a and 1b). Their methods often produce similar results, but they are based on different approaches. Both proceed by alternating steps in which the diffusion of heat is calculated with steps that compute the motion of the crystal surface. The main difference is in how they go about the second part of the calculation.

Almgren's approach treats the motion of the surface as a problem in geometric optimization. "At every step, you say 'What's the best shape that minimizes a certain energy function?' " he explains. Posing the problem in that way gives the approach an appealing conceptual generality. It also raises a number of theoretical questions and possibilities. In a good example of intergenerational as well as interdisciplinary research, Fred Almgren showed that the sequence of

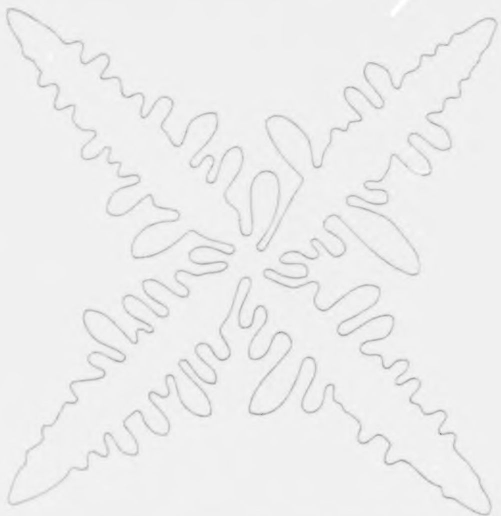


Figure 1a. *A crystal grown using Rob Almgren's approach.*

optimizing shapes do indeed approximate a smoothly growing crystal. In fact, this established the existence of a solution to the original problem, which had not been known till then. The theoretical results "might never have been obtained without the new ideas generated by trying to do numerical computation," Rob Almgren notes. Refining the results "remains an area of active research," he adds.

Roosen uses the same algorithm as Almgren for computing heat release and diffusion, but his approach is otherwise quite different. For one thing, Roosen works with surface energy functions that are associated with "completely facetted" interfaces, meaning that the boundary of the crystal is a polygon with sides set at prescribed angles (Almgren works with smooth energy functions, for which the interface is always a smooth curve or surface). His approach is also "more direct" than Almgren's. "What I do is say, 'At this point, how does [the crystal] move?' And then I move it. What [Almgren] does is say, 'How does the whole thing move?' "

Roosen's crystals grow in a five-step process. The key step comes first: Each edge of the interface is moved according to a rule that depends on the temperature along the edge and the crystal's "weighted mean curvature" at the edge — a concept introduced by Taylor to make sense of curvature in a setting where the curves consist of straight line segments at prescribed angles (as dictated by the anisotropic surface energy). Taylor (who is Roosen's thesis advisor) has developed much of the theory that establishes motion by weighted mean curvature as a practical approach to computational crystal growth.

The second step in Roosen's algorithm is a merging process in which, for example, edges that have shrunk to zero length are removed from the program's bookkeeping system. Next comes a "shattering" step which takes into account the fact that some parts of an edge may actually want to move faster than other parts because of an uneven temperature distribution. In the final two steps, the program computes the release of latent heat and its diffusion. These five steps are repeated tens of thousands of times. A typical calculation, Roosen says, takes four to ten hours.

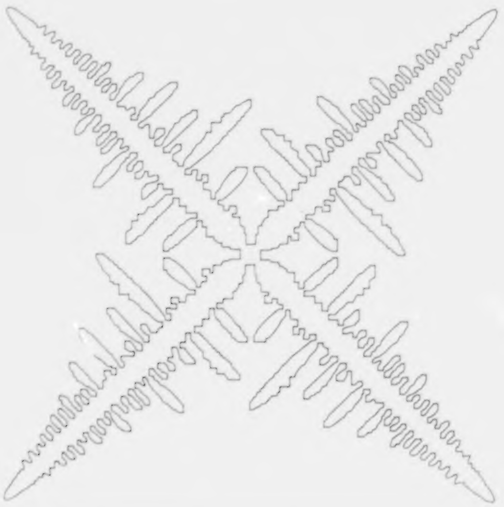


Figure 1b. *A crystal grown using Roosen's approach.*

That may seem like a long time to wait for a picture of a dendrite, but it's small potatoes compared to the run time for a full-blown three-dimensional computation. Both Almgren and Roosen's algorithms are conceptually suitable for three dimensions (much of the subject carries over easily into higher dimensions, for that matter), but the computational "load" can increase by a factor of several hundred. It would take weeks to simulate a single snowflake. It would take centuries to explore the variety produced in a single night's snowfall.

Improvements in the algorithms, further theoretical analysis, and more workstation horsepower are likely to bring 3-D calculations into the realm of practicality. However, another problem will remain: Figuring out a good way to visualize the results. Two-dimensional objects are easy to represent on paper or a computer screen; 3-D objects—especially objects you want to be able to see inside of—are far more challenging to represent.

Fortunately, the same issue crops up in a vast number of other problems, so a lot of thought has gone into this area. Researchers have made tremendous progress in recent years developing graphics programs that convert the computer's internal "knowledge" of an object into convincing, almost tangible pictures. It's something to look forward to: a 3-D movie (colorized, of course) of a computer-grown snowflake, surrounded by a glowing cloud of diffusing vapor.

The question is, will two such movies ever be alike?

A Growing Domain

Computational crystal growth is a wide-open field. Dendritic growth is only one of many open problems. Elizabeth Holm, a recent Ph.D. in materials science and scientific computation at the University of Michigan in Ann Arbor, is working on another: the structural evolution of "cellular arrays," such as occur in polycrystalline materials and—although it may seem far removed from the world of crystals—the foamy "head" on a glass of beer (see Figure 2).

As the name implies, polycrystalline materials are materials composed of many crystals, much as soap froth is composed of many individual bubbles. Over the course of time some domains grow while others shrink and disappear. The macroscopic properties of the material depend, in part, on the distribution of "grain" size in the crystalline microstructure.

Holm studies the process of domain growth with a computational model taken from statistical physics. The Potts model, as the approach is called, is like a "bitmap" of the microstructure, Holm explains. It describes the state of the material by an array of numerical indices assigned to a grid. In this setting, each domain consists of a contiguous set of grid sites that are assigned the same index. (On the computer screen, the indices are converted into colors.)

The evolution of the structure is modeled by any of a myriad set of rules. In one such rule, a single step of the algorithm is to pick a grid site at random, determine the number of neighboring grid sites with different indices, and if this number can be reduced by changing the index at the chosen site, then do so, otherwise either leave it alone or change it at random with some small, "temperature"-dependent probability.

Part of Holm's work has aimed at overcoming the effects of preferred directions (technically called "anisotropy") caused by the geometry of the grid. For example, in a square grid the boundaries between domains tend to be horizontal or vertical rather than diagonal. Holm and her coworkers have found that this inherent problem can be overcome in two ways: by extending the definition of "neighboring" grid site to a larger region (which, of course, entails more computation) or by increasing the "temperature" at which the simulation is performed. Their computations, including some 3-D simulations, indicate the Potts model should be useful for studying domain growth in a variety of physical systems.

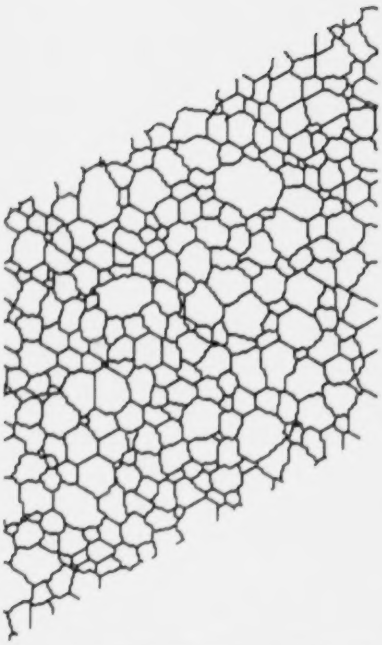


Figure 2. *Snapshot from Holm's computer study of the evolution of the grain microstructure in a two-phase polycrystal. (Figure courtesy of Elizabeth A. Holm, 1992.)*

BEST COPY AVAILABLE

Camp Geometry

You might call Kate Jenkins a budding mathematician. The Stanford sophomore spent last summer working on a computer program that parlays mathematical rules into pictures of flowering plants. Her geometrical bushes branch, bud, and even bend in a blowing breeze (see Figure 1).

Jenkins was one of nineteen undergraduates who participated in a summer research program at the Geometry Center in Minneapolis, Minnesota. The Geometry Center is a National Science Foundation (NSF) Science and Technology Center devoted to research at the cutting edge of geometry and computer visualization of geometric structures. But research is just one side of the coin: the Geometry Center also takes a serious interest in mathematics education.

Summer research programs for undergraduates in the mathematical sciences have become popular in recent years. The NSF last year awarded twenty grants in its Research Experience for Undergraduates program, at schools ranging from Williams College to Oregon State University. Many other colleges offered their own programs, as did research centers such as Los Alamos National Laboratory, the National Center for Atmospheric Research, and the Cornell National Supercomputer Facility.

"The philosophy is to give students a different experience with mathematics than the normal exam-packed classroom experience they get in school," says Al Marden, a professor at the University of Minnesota and director of the Geometry Center. The summer program gives students "a much more hands-on experience in mathematics, by doing it rather than by listening to somebody talking about it."

"It's sort of like an intellectual summer camp," adds Tony Phillips of the State University of New York at Stony Brook, who "coached" the students at the Geometry Center. For nine or ten weeks the students spent "all day and often part of the night" at the Center working on projects of their own choosing — "whatever they can think of," Phillips says.

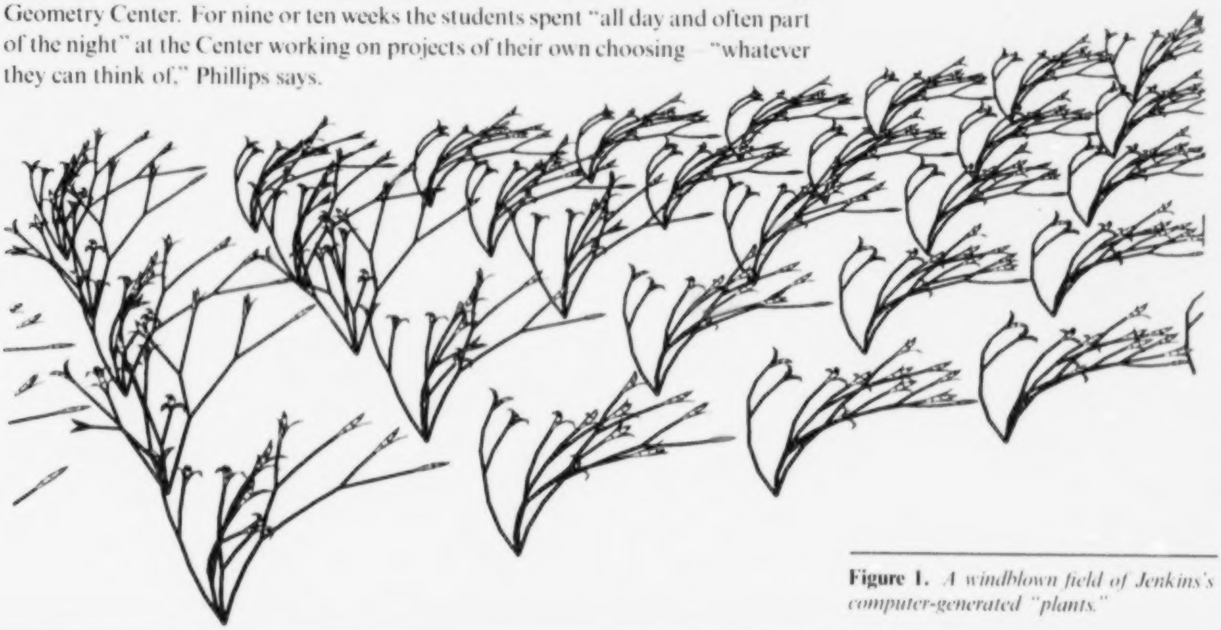


Figure 1. *A windblown field of Jenkins's computer-generated "plants."*

The summer program gives students "a much more hands-on experience in mathematics, by doing it rather than by listening to somebody talking about it," says Al Marden.

TH ETRY CENTER

1991 Summer Institute participants. Center: *John Hubbard (Cornell University).* Front row, seated: *Stephanie Mason, Carol Sohn, Albert Marden (Director of the Geometry Center), Anthony Phillips (Head Coach, SUNY, Stony Brook), David Broman, Jennifer Ellison.* Second row, standing: *David Ben-Zvi, Mark Meloon, Adrian Mariano, Sherry Scott, Gary Gutman.* Back row, standing: *Jacques Friedman, Karen Olsson, Craig Sutton, Linus Upson, Nicholas Coult, Kate Jenkins, Ken Bromberg. (Institute participants not in photo: Chris Cianflone, Thomas Colthurst, and Prem Janardhan.) (Photo by Chris Faust, Space Science Graphics, University of Minnesota.)*

Jenkins took her cue from the book *The Algorithmic Beauty of Plants* by Przemyslaw Prusinkiewicz and Aristid Lindenmayer. She wrote a computer program that draws plants using "L-systems"—instruction sets that create complex forms by the recursive application of simple rules. L-systems were introduced in theoretical biology by Lindenmayer in the late 1960s.

L-systems give geometric life to an otherwise abstract algebra of symbolic manipulations. For example, an L-system might start with the character string FFRF, which it interprets as Move Forward, Move Forward, Turn Right, Move Forward. The key ingredient is a replacement rule which turns each instance of each character in the "instruction string" into some other string of instructions. In some systems, a single character—usually the "Forward" instruction—is replaced by the *entire* original instruction set. Thus, for example, FFRF becomes (FFRF)(FFRF)R(FFRF), or, removing parentheses, FFRFFFRFRFFRF. If the replacement rule is applied several times and then the resulting instruction string plotted by, say, drawing a line segment with each forward move, the result can be an elaborate, even organic-looking picture.

Jenkins employed more complicated branching and growth rules to produce animated "cartoons" of developing plants. Using a dash of vector geometry, she also worked out ways for her plants to rustle in a simulated breeze and dip due to gravity.

Stephanie Mason, a junior at Virginia Tech, also worked with L-systems, but toward a totally different end: creating music. Mason takes the geometric result of an L-system and interprets it musically. For example, a vertical move may correspond to a step up or down in pitch, while horizontal moves indicate the duration of a note. The basic instruction set establishes a motif which iterations elaborate upon, she explains.

"You can actually create exactly what you want from these L-systems," Mason says.

One of her creations comes from an L-system that leads to a space-filling curve called the quadratic Gosper curve. Mason has set up the program so that the curve — really a set of line segments with right-angle turns — is drawn on a computer screen as the music is played on a synthesizer. Mason worked closely with Chris Cianflone, a student at the University of Minnesota (now in graduate school at the University of California at Berkeley), who developed an experimental musical program based on Fourier analysis of existing melodies.

Composers have long played with the formal structure of music. Bach, for example, is well known for writing music that could be played backward as well as forward. Mason has gone a step further, with music that can be played sideways as well, in what she calls a "right-angle canon." To do this, she simply takes a curve and rotates it so that pitch and duration are interchanged. When both curves are played together, using separate synthetic "voices" (Mason leans to piano and flute), the effect is surprisingly musical (see Figure 2).

"Bach would have loved it," Phillips remarks.

While Mason and Cianflone were turning Bach inside out, Nick Coult, a senior at Carleton College in Northfield, Minnesota, was putting a spring in orbit and numerically tracking the resulting motion. The idea, Coult says, was suggested by John Hubbard, a professor at Cornell who is on the permanent faculty at the Geometry Center. The problem is a variant on the three-body problem: There are

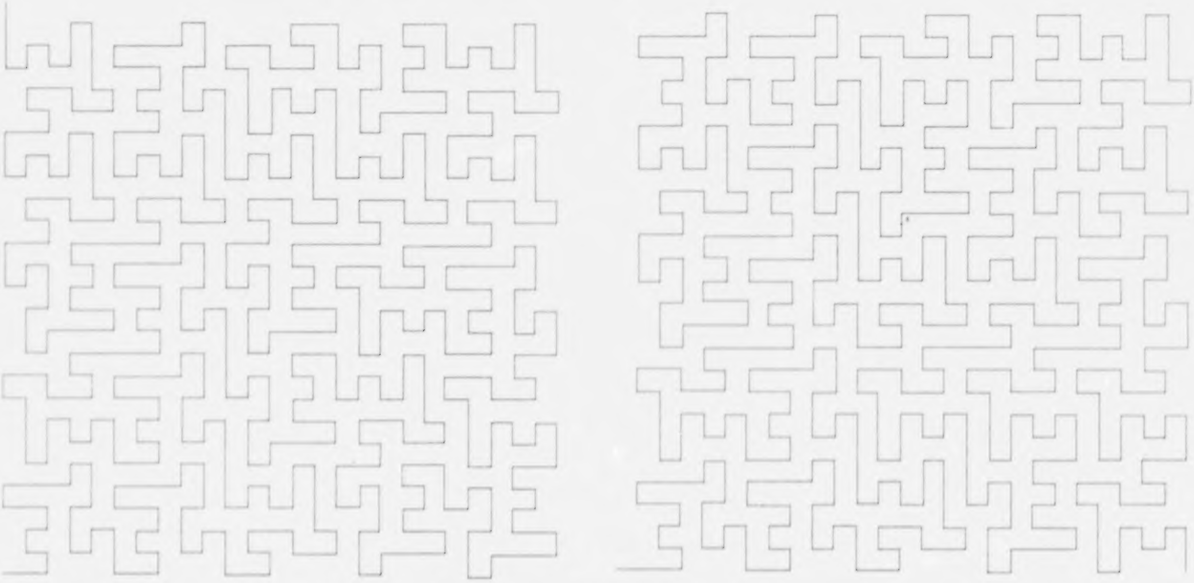


Figure 2. *Two examples of "musical scores" produced by Mason's computer program based on L-systems. The "flute" is on the left, the "piano" on the right. The flute starts in the lower left-hand corner, the piano in the lower right-hand corner.*

small masses at each end of the spring, and a large, gravitationally attracting mass which the mass-spring system orbits.

Hubbard had proposed the problem as a simplified model to test the hypothesis that tidal forces are responsible for the fact that we always see just one face of the moon. In this model, the earth is represented by the large mass and the moon is represented by the mass-spring system.

Coult indeed found that, in many cases, the system evolved into just such a stable orbit, with one end of the spring always pointing toward the central mass. However, he also found other stable orbits, some of which are surprisingly complex. "They're quite interesting to look at," Coult observes.

Setting up the equations that describe the motion was easy enough, Coult recalls. The challenge lay in solving them numerically, and then analyzing the solutions. For one thing, he says, "you have to verify that what you have on the computer screen [when it plots an orbit] is actually right." The programs he developed, Coult thinks, might be useful educational tools for courses in differential equations, although "that's not something I was thinking of when I started."

That sort of unexpected development is one of the benefits of letting students loose to do what they want in a relaxed atmosphere outside the usual classroom setting, Phillips points out. "Here, there's no test and there's no competition," he says. "The students are free to work at their own pace. They all work pretty hard, though."

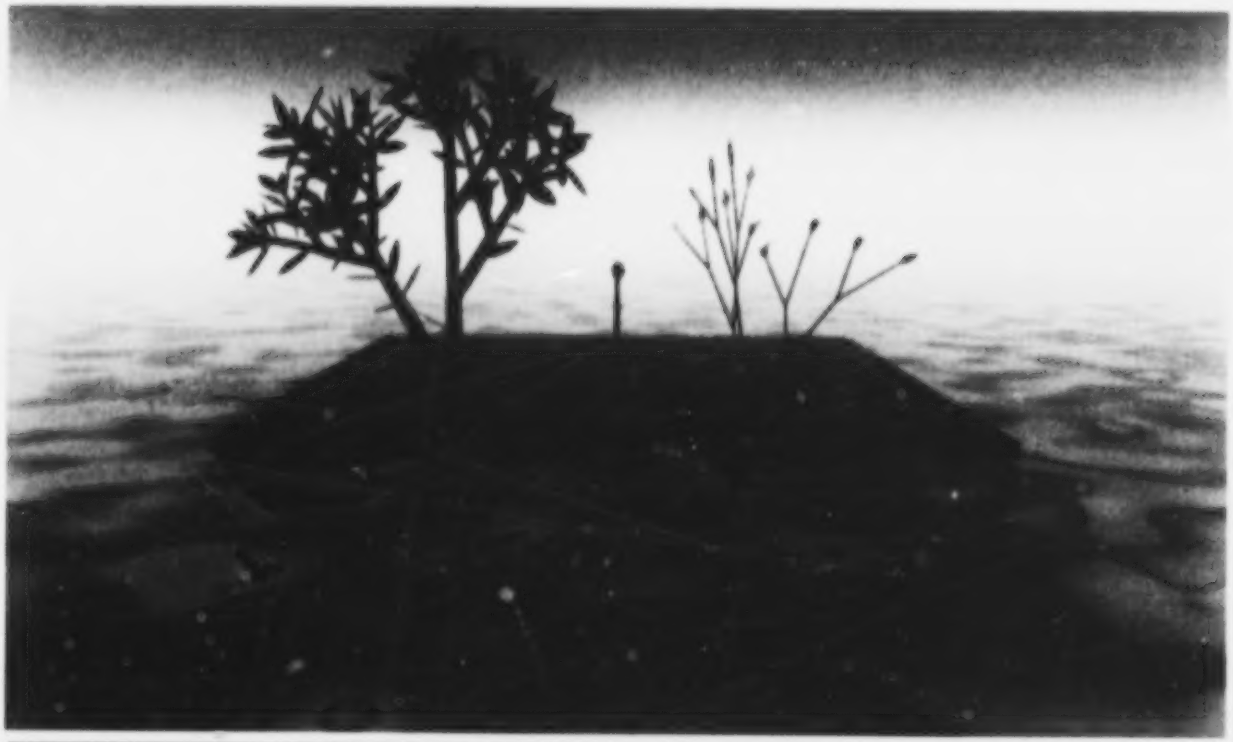


Figure 3. *Another example of Jenkins's computer-generated "gardens."*

Number Theorists Uncover a Slew of Prime Impostors

Finding large primes and factoring their composite progeny have turned out to have applications in such areas as cryptography and computer security systems.

One of the oldest and best known examples of mathematical reasoning is Euclid's proof that the sequence of positive integers contains infinitely many primes. Three mathematicians at the University of Georgia have recently put a curious — and potentially important — twist on Euclid's famous theorem. W. R. "Red" Alford, Andrew Granville, and Carl Pomerance have shown that there are also infinitely many prime *impostors*.

No doubt that calls for an explanation.

A prime number, of course, is a number that's divisible only by itself and 1. Primes have lots of desirable properties; they are often called the building blocks of number theory. However, unlike building blocks, where it's easy to tell the difference between a block and a building, it's not always obvious whether a number — say 12345678910111213 — is prime or composite (that is, a product of several primes).

At one time, the study of methods for identifying primes was considered an esoteric pursuit, even by many mathematicians. No more. Finding large primes and factoring their composite progeny have turned out to have applications in such areas as cryptography and computer security systems. The number-theoretic pursuit of efficient algorithms for primality testing and factorization also serves as a springboard for general ideas aimed at improving the efficiency of computer algorithms for other problems.

Computational efficiency is at the heart of the problem. If you're not worried about how long it takes, the definition of a prime gives you a simple way to tell if you've got one: Given a number like 12345678910111213, just try dividing it by 2, 3, 4, and on up to 12345678910111212. If one of these numbers divides 12345678910111213, then it's composite; if none does, then it's a prime. (Actually, it's only necessary to trial-divide up to the square root of the number in question, since the divisors of a number n can't all be greater than $\sqrt{n}$. Also, it's actually only necessary to trial-divide by *primes* up to the square root, since, for example, 6 can't divide a number if 2 and 3 didn't already divide it.)

Trial division works well when the number in question is small. But it's not a sensible way to verify the primality of large numbers. That's because the amount of computation it calls for gets quickly out of hand. For numbers with even just a few dozen digits, the computer run-times for trial-division primality testing start being measured in terms of the age of the universe.

Nobody wants to wait that long for an answer. It's like being put on hold when you're calling long distance.

But what can you do? Well, in 1640 — long before the lightning-fast computers of today — the French mathematician Pierre de Fermat discovered a property of prime numbers that provides a surprisingly efficient test for primality — usually. What Fermat found is the following: If n is a prime number and a is *any* number whatsoever, then the number $a^n - a$ is divisible by n.

That statement has come to be called Fermat's Little Theorem (as distinct from the more famous "Last Theorem"). In spite of the diminutive title, Fermat's Little

Theorem is one of the most important results in number theory. In a very real sense it's the cornerstone of the subject.

For the purpose of primality testing, though, the value of Fermat's Little Theorem is in its ability to expose numbers that are *not* prime. It does this by turning the statement around: If n is a number and a is some number such that $a^n - a$ is *not* divisible by n, then n very definitely is *not* a prime number.

This turns out to be a remarkably efficient test: Given n, see if it divides $2^n - 2$, $3^n - 3$, $5^n - 5$ and maybe a few others (once again, it's only necessary to try prime numbers for a); if it fails *even once*, then n is not prime. For example, Fermat's Little Theorem "proves" that 6 is composite (if that were ever in doubt!)

The chance of being burned by a prime impostor is pretty low—legitimate primes are far more common than Carmichael numbers—but it's still worth reading the results with a *caveat calculator*.

COMPOSITE
WHICH ALWAYS
DIVIDE a^n-a

Putting their heads together to solve a tough number theory problem: Andrew Granville (top), Red Alford (left), and Carl Pomerance. (Photo by Rick O'Quinn, University of Georgia News Bureau.)

43 BEST COPY AVAILABLE

because 6 fails to divide $2^6 - 2 = 62$. The fact that 6 does divide $3^6 - 3 = 726$ makes no difference; for a number to be prime, it must *always* pass the test imposed by Fermat's Little Theorem.

What makes this a good test for primality is that composite numbers tend to be exposed very quickly. Most of the time it's only necessary to test $2^n - 2$ — the smallest composite number that slips by that test is $n = 341 = 11 \times 31$. It's very rare for a composite number to pass Fermat's test for more than a couple of values of a.

But it does happen. In fact, there are composite numbers that pass Fermat's test for *all* values of a. The smallest such number is $561 = 3 \times 11 \times 17$. Then come 1105, 1729, 2465, and 2821. These numbers are impostors; they "masquerade as prime numbers," says Granville. He calls them "annoying."

Infinitely annoying.

The first examples of these prime impostors were found around 1910 by American mathematician Robert D. Carmichael, and for that reason they are called "Carmichael numbers." It would be one thing if there were just a few of them — that would offer hope that Fermat's Little Theorem could be used not just as a proof of compositeness, but also as a guarantee of primality. But theorists kept finding more of them, and it seemed likely the list would prove endless.

That's exactly what Alford, Granville, and Pomerance have now shown: There are infinitely many Carmichael numbers. Moreover, their work makes it clear that many, if not all, other primality tests based on ideas similar to Fermat's Little Theorem are equally flawed by infinite families of composite numbers that pass the various tests. "There's no way to just generalize Fermat's Little Theorem to a [perfectly accurate] primality test," says Alford — or rather, he adds, "there's probably no simple way" to do it.

That may be news to users of computer algebra systems such as *Mathematica*. These systems, which manipulate symbolic expressions and do "exact" arithmetic, generally include a primality test based on one of the jazzed-up versions of Fermat's Little Theorem. Again, when one of these tests says a hundred- or thousand-digit number is composite, the result is reliable (even though, paradoxically, the test doesn't contain any clue as to what the factors might be!), but when the test says "prime," it really means "probably prime."

"It's not generally appreciated that these tests are not proofs of primality," says Pomerance. The chance of being burned by a prime impostor is pretty low — legitimate primes are far more common than Carmichael numbers — but it's still worth reading the results with a *caveat calculator*.

The Georgia trio's proof that Carmichael numbers pop up infinitely often is based on a heuristic argument put forward by Paul Erdős in 1956. The main idea is to choose a number L for which there are a large number of primes p that don't themselves divide L, but have the property that $p - 1$ divides L. The key point is then to show that these primes can be multiplied together in lots of different ways so that the products all leave remainder 1 when divided by L. It turns out that each such product is a Carmichael number.

For example, with $L = 120$, the primes in question are 7, 11, 13, 31, 41, and 61. A check of all possible combinations reveals that $41041 = 7 \times 11 \times 13 \times 41$, $172081 = 7 \times 13 \times 31 \times 61$, and $852841 = 11 \times 31 \times 41 \times 61$ all leave remainder 1 when divided by 120, and hence are Carmichael numbers.

The fact that the numbers constructed by Erdős's argument are Carmichael

"And then here I was, I had done all of this thinking about how to do it. . . and I was crushed, I was really crushed," says Alford. "So I went home that night, really with my nose just plain flat out of joint. But the next morning I woke up, and I knew how to construct 2^{100} of them."

numbers goes back to the mathematician A. Korselt. Korselt proved that a number n divides all numbers of the form $a^n - a$ (that is, n is a Carmichael number) if and only if it is squarefree (which means no prime divides it more than once) and has the property that $p - 1$ divides $n - 1$ whenever p is a prime divisor of n. Interestingly, he proved this in 1899—more than a decade before Carmichael put his stamp on the subject. The difference is, Korselt seemed to think there were no such numbers. They "surely would have been known as Korselt numbers had he just done a few computations!" says Granville.

What Alford, Granville, and Pomerance did, in essence, was to make Erdős's argument precise. Their work was spurred by recent work of Zhang Mingzhi at Sichuan University, who used the Erdős heuristic to produce examples of large Carmichael numbers. Alford, who is mainly a computational number theorist, thought he could do better. "I was champing at the bit to construct one a million digits long," he recalls. Instead, Pomerance challenged him to show that his method would produce huge families of such numbers.

"He said, 'Red, if it's as easy as you say it is, why don't you see if you can't construct 2^{50} of them.' " Alford recalls. "And then here I was, I had done all of this thinking about how to do it. . . and I was crushed, I was really crushed, for Carl to say it's just like writing down a bigger integer!" Alford laughs, then continues: "So I went home that night, really with my nose just plain flat out of joint. But the next morning I woke up, and I knew how to construct 2^{100} of them."

In fact, Alford's program coughed up 2^{128} Carmichael numbers (or, rather, prime numbers that can be combined in that many ways to produce Carmichael numbers). That enticed Granville and Pomerance to tackle the theoretic end of Erdős's argument. By using deep results in analytic number theory and combinatorial techniques from the theory of groups, they were finally able to flesh out Erdős's argument enough to prove a succinct theorem: There are more than $x^{2/7}$ Carmichael numbers up to x, for all sufficiently large x.

Exactly how large x has to be to be "sufficiently large" is still unclear (the analytic techniques are too convoluted to produce an estimate), although the numerical evidence suggests it happens at around $x = 10^7$. In a way it doesn't matter much, because Erdős's argument actually implies that the exponent 2/7 can be replaced by any value short of 1. In other words, if you go far enough out, Carmichael numbers are amazingly abundant. The impostors are not as numerous as the true primes, but there are enough of them to make you stop and wonder.

An Imposing New Prime

In the early months of 1992—about the same time the Georgia trio proved the infinitude of prime impostors—a group at AEA Technology's Harwell Laboratory in Britain announced the discovery of a new "largest" prime number: a monster with nearly a quarter of a million digits belonging to a class of numbers known as "Mersenne primes."

The new number is not, of course, the largest possible prime. There is no such thing. What "largest" means here is it's the largest number *known* to be prime.

Mersenne primes all have the form $2^p - 1$, where the exponent p is itself a prime. If all primes p produced Mersenne primes, there'd be no sport in finding "largest" primes, but that's not the case. Mersenne primes are rare enough that only thirty-two of them are known to exist. The one found last year occurs for the exponent $p = 756{,}839$. A Cray-2 supercomputer took nineteen hours, using a variant of Fermat's Little Theorem known as Lucas's Test, to verify that the number indeed is prime.

Map-Coloring Theorists Look at New Worlds

People often think that once a "hard" mathematical problem has been solved, that's the end of the story. Nothing could be further from the truth. Mathematical problems rarely exist in a vacuum; the best ones are usually surrounded by a coterie of other interesting problems. Rather than spelling the end to a subject, the solution to a challenging problem more often means that researchers will turn with increased interest to some of the questions related to it.

Take the Four Color Theorem, for example. In 1976, Kenneth Appel and Wolfgang Haken at the University of Illinois proved that four colors are enough to paint any conceivable map in the plane in such a way that no countries with a common border are painted the same color. This breakthrough, which the researchers summarized as "four colors suffice," was one of the most talked-about results of the 1970s, in part because much of the proof was done on a computer. To many, the story of map coloring seemed over and done with.

Not so. Recently, researchers have been looking at the map-coloring problem for classes of maps to which the Four Color Theorem does not apply. These maps are drawn not on a flat piece of paper, but on arbitrary surfaces with any number of "handles" on them, such as a coffee cup or a two-handled vase. What the researchers have found can be stated as a nice counterpart to Appel and Haken's result: For these new classes of maps, *five* colors suffice.

Actually, there are two separate five-color theorems. Carsten Thomassen, a mathematician at the Technical University of Denmark, has proved that five colors are enough for maps on these many-handled surfaces, provided the countries to be colored are sufficiently small and numerous. Meanwhile, Neil Robertson at Ohio State University, Paul Seymour at Bell Communications Research (Bellcore) in Morristown, New Jersey, and Robin Thomas at Georgia Institute of Technology have reached the same conclusion for a different class of maps: Five colors suffice provided the countries to be colored can't be aligned into six mutually neighboring "federations."

The provisos of the two five-coloring theorems are poles apart, Robertson points out. That doesn't mean the two theorems are in conflict, though. Quite the contrary, it means the results combine to account for a large class of maps drawn on general surfaces.

Both new theorems belong to a branch of mathematics called graph theory. A mathematical graph is an extremely simple object: just a bunch of points (called *vertices*) with a bunch of curves (called *edges*) connecting them. Few things are simpler than that, yet few things lead so quickly to complicated problems and intricate results. Graph theory has come to play a key role in theoretical computer science and numerous other applied areas ranging from the design of transportation networks to the mathematics of chemical compounds (see box on page 46). Map coloring is just one area where graph theory plays a unifying role.

It's very easy to turn a map into a graph. You put a vertex at the capital of each country, and draw an edge between two vertices if the corresponding countries have a common border (see Figure 1). The graphs that result from ordinary maps

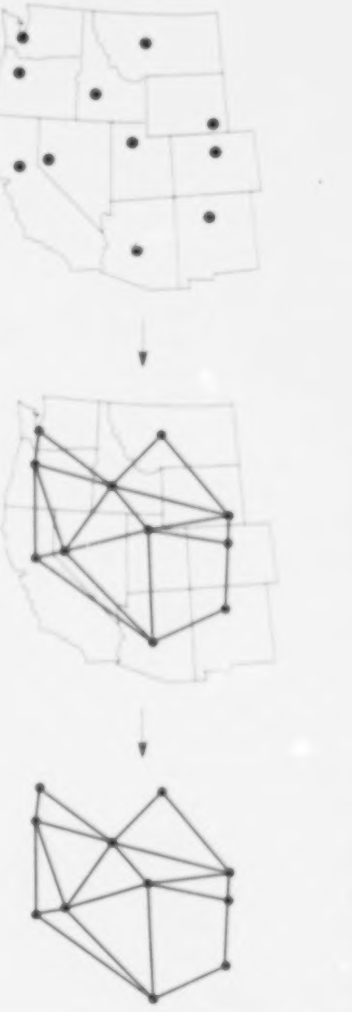


Figure 1. *Any map on the plane can be converted into a planar graph, and vice versa. Here, state capitals are connected by lines to form a graph.*

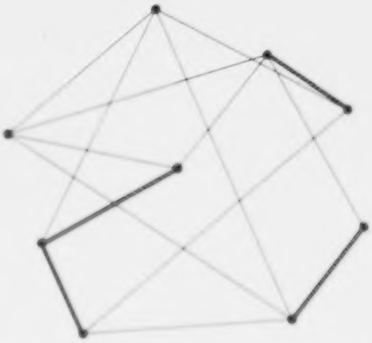


Figure 2. *If the dark edges are contracted so that their vertices merge, the resulting graph (called a "minor") is identical to K_5. Therefore this graph cannot be drawn on the plane without at least two edges crossing.*

are called "planar" graphs, because they can be drawn on a flat plane without any of their edges crossing. Because each vertex corresponds to a country, "coloring" a graph means coloring the vertices: if two vertices are connected by an edge, they should be given different colors. The idea that the edges do not cross corresponds to the fact that if two countries have a common border, you can make a road between the two capitals that travels only in those two countries.

The Four Color Theorem says that you need at most four different colors to color any planar graph. But not every graph is planar. In particular, the graph with five mutually adjacent vertices cannot be drawn in the plane without two edges crossing. (If it could, the Four Color Theorem would be false!) Another example is the "bipartite" graph consisting of two groups of three vertices, with an edge connecting each vertex in one group to each vertex in the other.

In 1930, the Polish mathematician Kazimir Kuratowski proved that these two graphs (usually denoted as K_5 and $K_{3,3}$) are essentially the only nonplanar graphs. What this means is that any other nonplanar graph contains at least one of these two, possibly in the form of a "minor," which is the graph theorist's term for a set of federations. (A federation forms when two neighboring countries erase their common border. For graphs, this amounts to contracting an edge until the vertices it connects merge. See Figure 2.) In other words, there are essentially only two "obstructions" to a graph being planar: K_5 and $K_{3,3}$.

While K_5 and $K_{3,3}$ cannot be drawn on a flat piece of paper, they can be drawn on a surface with a handle (see Figure 3). And indeed, every graph can be drawn on some surface with some number of handles, although determining exactly how many handles are necessary is not easy in 1988 Thomassen proved that task to be NP-complete. (For an explanation of NP-complete problems, see "New Computer Insights from 'Transparent' Proofs," pages 7-11.)

In 1983, Robertson and Seymour generalized Kuratowski's theorem from the plane to surfaces with handles. They proved that for each such surface the set of "obstructing" graphs, while possibly quite large, is nevertheless always finite.

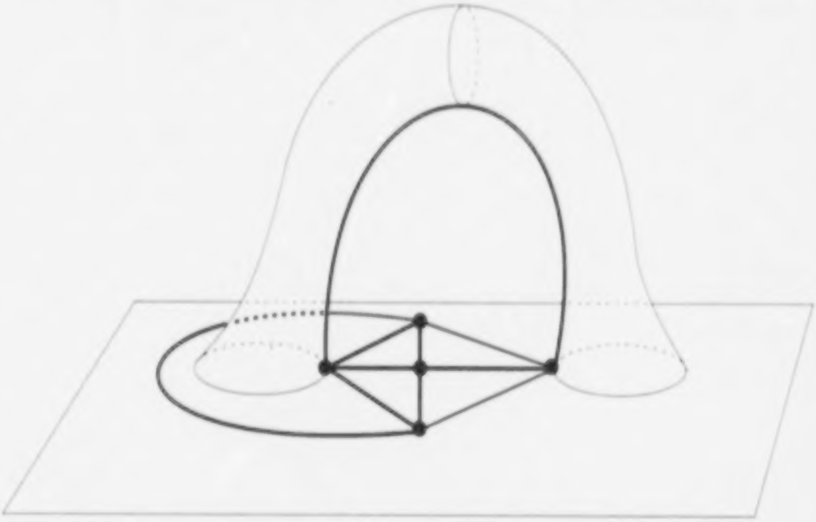


Figure 3. *K_5 can't be drawn on a plane without at least two edges crossing. But it can be drawn on a surface with a handle.*

Five Colors *Don't* Suffice!

In 1890, Percy Heawood proved that any map drawn on flat paper can be colored with no more than five different colors. Heawood's five-color theorem also applies to maps drawn on the globe, because, topologically speaking, the sphere and the plane are equivalent. However, it doesn't hold for graphs drawn on surfaces with handles. For example, it's possible to divide the torus into seven regions, each of which borders the other six, thus necessitating a separate color for each region.

How many colors does a surface with handles require? Heawood proved it never takes more than $(7 + \sqrt{48g + 1})/2$ colors, where g is the number of handles. In 1968, Gerhardt Ringel and J. W. T. Youngs proved that Heawood's bound (rounded down to the nearest integer) is exact: Every surface has maps that require as many colors as allowed by the formula.

Curiously, Heawood's formula gives the right answer—four—for the plane ($g = 0$), but the proof only works when the surface actually has handles. This is just one of many instances in mathematics where a problem is easier to solve in a complicated-sounding setting than it is in its original guise.

The Four Color Theorem says that you need at most four different colors to color every planar graph. But not every graph is planar.

This was an early result in an ongoing research program aimed at developing a "structural" theory of graphs.

Their latest work with Thomas is in the same vein. In 1943, Hugo Hadwiger conjectured that any graph can be colored with n colors provided it doesn't contain K_{n+1} (the analog of K_5, with $n + 1$ mutually adjacent vertices) as a minor. (Obviously if a graph contains K_{n+1} outright, there's no way to color it with just n colors. If K_{n+1} is present as a minor, it may still be possible to get by with n or even fewer colors, as in Figure 4, but that's not what Hadwiger's conjecture is concerned with. His conjecture is concerned with those graphs that don't contain K_{n+1} in any way, shape, or form.) In other words, Hadwiger's conjecture says the only potential obstruction to n-coloring a graph is the presence of K_{n+1}.

Hadwiger's conjecture is clearly true for $n = 1$ and 2, and easy to prove for $n = 3$. For $n = 4$, it turns out to be equivalent to the Four Color Theorem (although the equivalence is by no means easy to prove). Robertson, Seymour, and Thomas's theorem settles the case $n = 5$. Of course, that leaves an infinite amount of Hadwiger's conjecture unsolved. But "it says we're beginning to learn what's going on," Robertson notes.

Hadwiger's conjecture really doesn't care much what kind of surface a graph or its associated map is drawn on. Thomassen's five-color theorem, however, does. The problem it solves has a somewhat shorter history. In 1982, Michael Albertson at Smith College in Northampton, Massachusetts, and Walt Stromquist, a mathematician at Dan Wagner Associates in Paoli, Pennsylvania, proved that any map drawn on a torus (that is, a surface with one handle, like a coffee cup or its topological cousin, the doughnut) can be five-colored provided that any "tour" that travels all the way around the handle in any direction visits at least eight different countries.

Albertson and Stromquist conjectured that something similar should be true for surfaces with more handles: that if all trips around the handles are sufficiently long (that is, visit sufficiently many countries), then the map (or graph) should be five-colorable. That condition is known as "local planarity."

At one point it was thought that locally planar maps might even be four-colorable. However, Steve Fisk at Bowdoin College in Brunswick, Maine, put an end to that in 1978, by showing how to draw maps on the torus (or any other

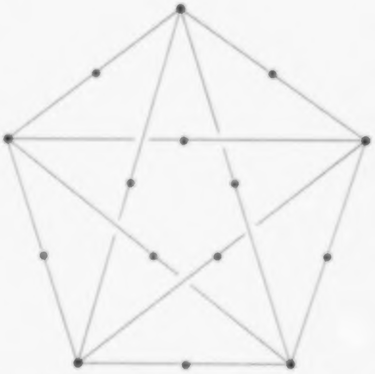


Figure 4. *This graph has been two-colored even though it contains K_5 as a minor. Does that violate Hadwiger's conjecture? No!*

surface with handles) with countries as small as you please that nevertheless require five colors.

In 1985, Joan Hutchinson, now at Macalester College in St. Paul, Minnesota, proved that five colors suffice if "small" is defined in a particular way. Thomassen's theorem goes further. It says that for each surface, there is a number n such that if all trips that go around handles visit at least n countries of a given map, then that map is five-colorable. The number n depends on the number of handles on the surface: Thomassen's proof provides only an estimate, which doubles each time another handle is added. The actual number, he notes, is likely to be a good deal less.

One may well ask, why bother? One answer is that graph coloring is not a purely academic exercise: it does have applications in real settings. "What graph coloring is really about is scheduling," says Stromquist. "Trying to color a graph is the same as trying to schedule a whole lot of events into time slots in such a way that incompatible events don't happen at the same time." Researchers interested in developing algorithms for scheduling consequently often find themselves faced with problems in graph coloring.

But the main reason continues to be one of pure intellectual challenge. The proof of the Four Color Theorem still requires a computer to rule out a thousand or so ways a planar map might require five colors rather than four, and to many mathematicians that's an unsatisfactory state of affairs. "There's obviously still something to be learned about graph coloring," says Robertson. Theorists believe that the results of research in map coloring will be of use in other areas of graph theory and its applications. Beyond that, there's one final, inarguable reason, summarized succinctly by Stromquist: "It's fun."

Fan Chung and Shlomo Sternberg with a model of a buckyball.

Graph Theory Tackles the Buckyball

Graph theory is one of the most playful topics in mathematics. But it's also one of the most useful. In part because of the way it combines algebraic abstractions with down-to-earth geometric configurations, graph theory turns up in places you might not expect—and graph theorists often wind up working in areas seemingly far afield.

Take Fan Chung, for example. Chung, a mathematician at Bellcore, is an expert on graph invariants. Normally she works either in pure theory or on applications to problems in communication networks. But recently she has been applying her graph-theoretic expertise to something quite different: buckyballs.

First discovered just a few years ago, the buckyball is a soccerball-shaped molecule consisting of sixty carbon atoms arranged in a highly symmetric, icosahedral pattern. An outline of the buckyball's chemical bonds is reminiscent of the "geodesic domes" popularized in the 1960s by R. Buckminster Fuller—hence the official chemical name, *buckminsterfullerene*.

Chemists, physicists, and materials scientists have flocked to the buckyball and its variants like moths drawn to candlelight (indeed, the soot that rises with the flame of a candle may consist in part of buckyballs). The crowd now includes a handful of mathematicians.

Graph theory is no newcomer to chemistry: The term "graph" was first used (in its technical sense) by the mathematician J. J. Sylvester in 1877, in a paper titled "Chemistry and Algebra." Mathematicians' drawings of graphs and chemists' renderings of chemical compounds are strikingly similar. That's no coincidence. Graph theory turns out to be a useful mathematical tool for chemists seeking to understand the myriad ways that elements can combine to form complex molecules and the myriad properties those molecules may possess. *(continued on next page)*

Chung has been working with Shlomo Sternberg at Harvard University to analyze the mathematical properties of the buckyball's unique structure. Sternberg is an expert in the theory of group representations, which can be roughly described as the systematic study of symmetry. The combination of graph theory and representation theory, applied to the buckyball, is a powerful one. Chung and Sternberg's analysis so far has gone a long way toward explaining the buckyball's spectroscopic properties and why the molecule is so stable.

Chung and Sternberg are also trying to find a theoretical explanation of another property of the buckyball, which it shares with a growing family of promising new materials, namely high-temperature superconductivity. The discovery of high-temperature superconductivity in the late 1980s has left theorists scrambling for explanations of a phenomenon that isn't even well understood for low temperatures. "It's really a very big puzzle," says Chung. However, the fact that so many materials turn out to be superconducting suggests that the puzzle may not depend on details of the physics. "It's our belief it must be a mathematical explanation," Chung concludes.

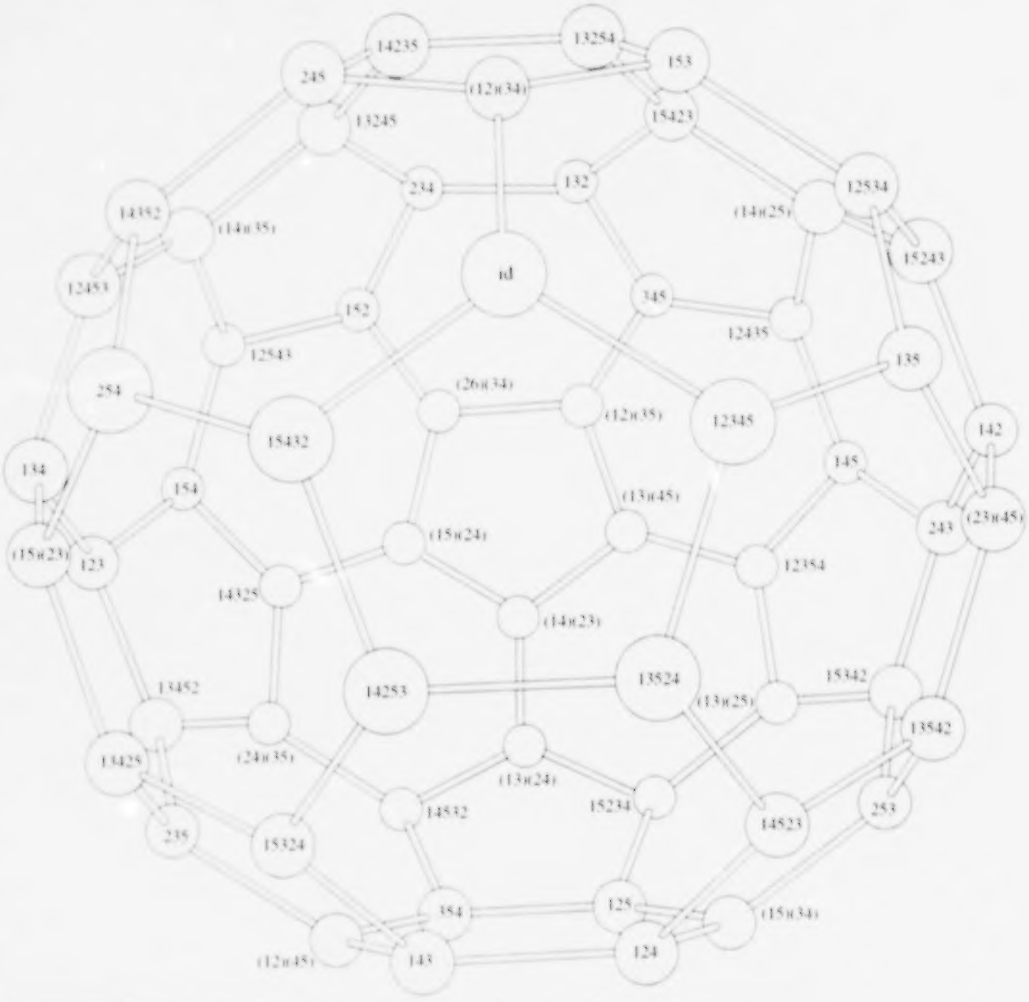

14235
13254
245
(12)(34)
153
13245
15423
234
132
14352
(14)(35)
12534
(14)(25)
12453
id
15243
152
345
12435
12543
135
254
(26)(34)
(12)(35)
12345
15432
142
145
134
154
(13)(45)
243
(23)(45)
(15)(23)
123
(15)(24)
12354
14325
(14)(23)
13452
14253
13524
15342
(13)(25)
13542
13425
(24)(35)
(13)(24)
235
14532
15234
14523
253
15324
354
125
(15)(34)
(12)(45)
143
124

Figure 5. *The Chung-Sternberg buckyball. (Based on figure courtesy of Fan R. K. Chung.*

Credits

ADVISORY BOARD

Fan R. K. Chung
Bell Communications Research

Avner Friedman
Institute for Mathematics and its Applications

James G. Glimm
State University of New York, Stony Brook

Benedict H. Gross
Harvard University

David O. Siegmund
Stanford University

William P. Thurston
Mathematical Sciences Research Institute

About the Author: Barry Cipra is a freelance mathematics writer based in Northfield, Minnesota. He is currently a Contributing Correspondent for *Science* magazine and also writes regularly for *SIAM News*, the newsletter of the Society for Industrial and Applied Mathematics. He received the 1991 Merten M. Hasse Prize from the Mathematical Association of America for an expository article on the Ising model, published in the December 1987 issue of the *American Mathematical Monthly*. His book, *Misteaks ... and how to find them before the teacher does...* (a calculus supplement), is published by Academic Press.

Project Administration: Samuel M. Rankin, III, AMS Associate Executive Director and Director of Publications

Editorial Direction: Allyn Jackson, Staff Writer

Production Coordination: Michael A. Kowalski, Production Editor

Typography: Neil G. Bartholomew, Technical Support;
Ralph E. Youngen, Technical Support Manager

Design: Peter B. Sykes, Staff Artist

The AMS gratefully acknowledges the support of the Exxon Education Foundation and the Alfred P. Sloan Foundation for the publication and distribution of *What's Happening in the Mathematical Sciences*.

About the American Mathematical Society

The American Mathematical Society is a nonprofit organization devoted to research in the mathematical sciences. For more than 100 years, the Society has worked to support the advance of mathematical research, the communication of mathematical ideas, and the improvement of the mathematics profession. In recent years, the Society has increased its attention to mathematics education, public awareness of mathematics, and the connections of mathematics research to its uses.

The AMS is the world's largest mathematical organization, with nearly 30,000 members. As one of the world's major publishers of mathematical literature, the Society produces a wide range of book series, journals, monographs, and videotapes, as well as the authoritative reference, *Mathematical Reviews*. The Society is a world leader in the use of computer technology in publishing and is involved in the development of electronic means of information delivery. Another primary activity of the AMS is organizing meetings and conferences. In addition to the annual winter meeting, the Society organizes bi-annual summer meetings, as well as numerous smaller meetings during the academic year and workshops, symposia, and institutes during the summer. Other major Society activities are employment services, collection of data about the mathematical community, and advocacy for the discipline and the profession.

The main headquarters of the Society, located in Providence, Rhode Island, employs nearly 200 people and contains a large computer system, a full publication facility, and a warehouse. Approximately eighty people are employed at the *Mathematical Reviews* office in Ann Arbor, Michigan. In September 1992 the AMS opened an office in Washington, DC, in order to enhance the Society's public awareness efforts and its linkages with federal science policy.

The American Mathematical Society is pleased to offer single issues of this publication free of charge. Shipping and handling is $7.00. For optional delivery by air to foreign addresses, please add $6.50. To order, please specify HAPPENING/1. Write to:

Membership and Customer Services
American Mathematical Society
P.O. Box 6248
Providence, RI 02940-6248

Telephone orders: To use Visa or MasterCard call 1-401-455-4000, or, in the U.S. and Canada, 1-800-321-4AMS (321-4267)

Because of limited quantities, requests for multiple copies will be considered on a case-by-case basis. Requests should be sent to the AMS Director of Publications, Samuel M. Rankin, III, AMS, P.O. Box 6248, Providence, RI 02940-6248

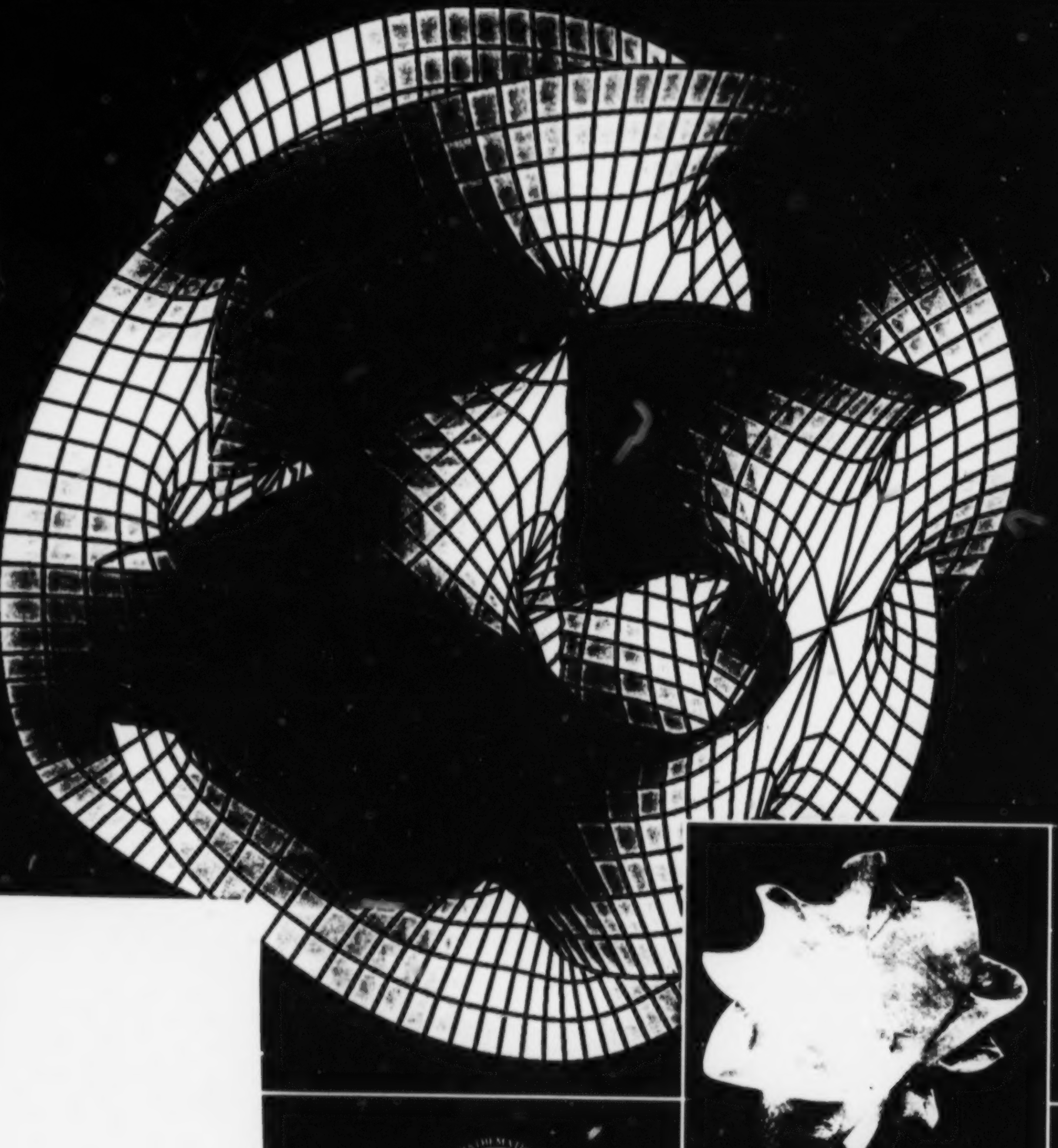
ISBN 0-8218-8999-0

9 780821 889992

BEST COPY AVAILABLE

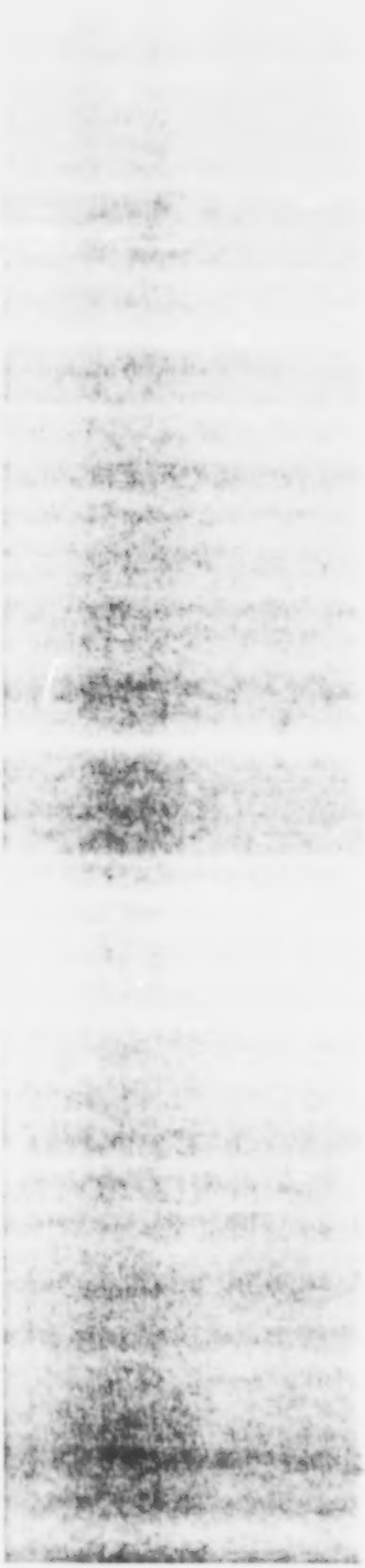
What's Happening in the Mathematical Sciences

Volume 2 • 1994



American Mathematical Society

BEST COPY AVAILABLE



Introduction

Welcome to the 1994 issue of *What's Happening in the Mathematical Sciences*, a yearly publication of the American Mathematical Society, inaugurated in 1993. Volume 2 continues the theme of surveying some of the important developments in the mathematical sciences over the past year or so. One purpose of *What's Happening* is to convey that mathematics is a dynamic discipline, contributing to research and development in many areas of science as well as contributing significantly to the solving of some of the major problems facing society. In this issue you can read about a mathematically-based technology that produces real time continuous images of the heart, lungs, and other organs; results on key problems in the area of knot theory and how these results lead to insights in the study of DNA; recent findings in the theory of waves; and Fermat's Last Theorem.

What's Happening in the Mathematical Sciences is written in a style so that the general public can learn about the beauty and universality of mathematics. The American Mathematical Society hopes you enjoy it.

Samuel M. Rankin, III

Samuel M. Rankin, III
AMS Associate Executive Director

Front Cover. A collaboration between computer scientist Andrew Hanson at Indiana University and artist Stewart Dickson in Los Angeles has brought the Fermat equation $x^n+y^n=z^n$ to life. The computer graphic shows a 3-dimensional projection of the complex Fermat surface $u^5+v^5=1$ (the exponent is indicated by the 5 grid lines that intersect at a point). Dickson has used a high-tech process called stereolithography to render the surface as a truly 3-dimensional sculpture.

Back Cover. New York-based sculptor Rhonda Roland Shearer combines elements of modern fractal geometry, expressed through plant forms, with classical Euclidean geometry in *The 5 Platonic Solids: Terra* (blue patina cube), *Ignis* (yellow ochre patina tetrahedron), *Aqua* (red patina dodecahedron), *Aer* (orange patina octahedron), and *Caelum* (viridian green patina icosahedron). (Photo courtesy of Lee Boltin. Copyright © 1992, by permission of Rhonda Roland Shearer.)

What's Happening in the Mathematical Sciences

Volume 2 • 1994

Written by Barry Cipra
Edited by Paul Zorn

Contents

"A Truly Remarkable Proof" 3
The announcement last year of a proof of Fermat's Last Theorem stunned the mathematical world. Andrew Wiles's proof, though currently incomplete, has nonetheless drawn rave reviews.

From Knot to Unknot 9
What's the quickest way to untie a knot? Researchers have untangled a good part of the answer.

New Wave Mathematics 14
Will compact waves cruise the information superhighways of the future? In theory, it's possible.

Mathematical Insights for Medical Imaging 19
A team of mathematicians, computer scientists, and engineers has designed a new medical imaging technology based on the safe application of electric currents.

***Parlez-vous* Wavelets?** 23
Mathematicians and scientists are rapidly learning to speak a new language. The results are making a big splash.

Random Algorithms Leave Little to Chance 27
Computer scientists will do anything to avoid bottlenecks and speed up computations. But gamble on the results? You bet!

Soap Solution 33
Undergraduate students at a summer mathematics research program have found some slick answers to some old problems about the geometry of soap bubbles.

Straightening Out Nonlinear Codes 37
A complicated class of error-correcting codes has suddenly gotten much easier to use.

Quite Easily Done 41
A combinatorial problem, long thought to be difficult, has finally been solved—with surprising ease.

(Vector) Field of Dreams 47
A clever construction "pulls the plug" on a 40-year-old conjecture about the topology of vector fields.

ISBN 0-8218-8998-2
ISSN 1065-9358

© 1994 by the American Mathematical Society.
All Rights Reserved.

Permission is granted to make and distribute verbatim copies of this publication or of individual items from this publication provided the copyright notice and this permission notice are preserved on all copies.

Permission is granted to copy and distribute modified versions of this publication or of individual items from this publication under the conditions for verbatim copying, provided that the entire resulting derived work is distributed under the terms of a permission notice identical to this one.

1991 *Mathematics Subject Classification*:
Primary 00A06

Printed in the United States of America.

This publication has been typeset using the TEX typesetting system running on a Solbourne 5/502 Unix computer. Halftones were created from original photographs with Adobe Photoshop, and illustrations were redrawn using Adobe Illustrator on Macintosh Quadra computers. PostScript code was generated using **dvips** by Radical Eye Software.

Typeset on an Agfa/Compugraphic 9600 laser imagesetter at the American Mathematical Society. Printed at E. A. Johnson, East Providence, RI, on recycled paper.

BEST COPY AVAILABLE

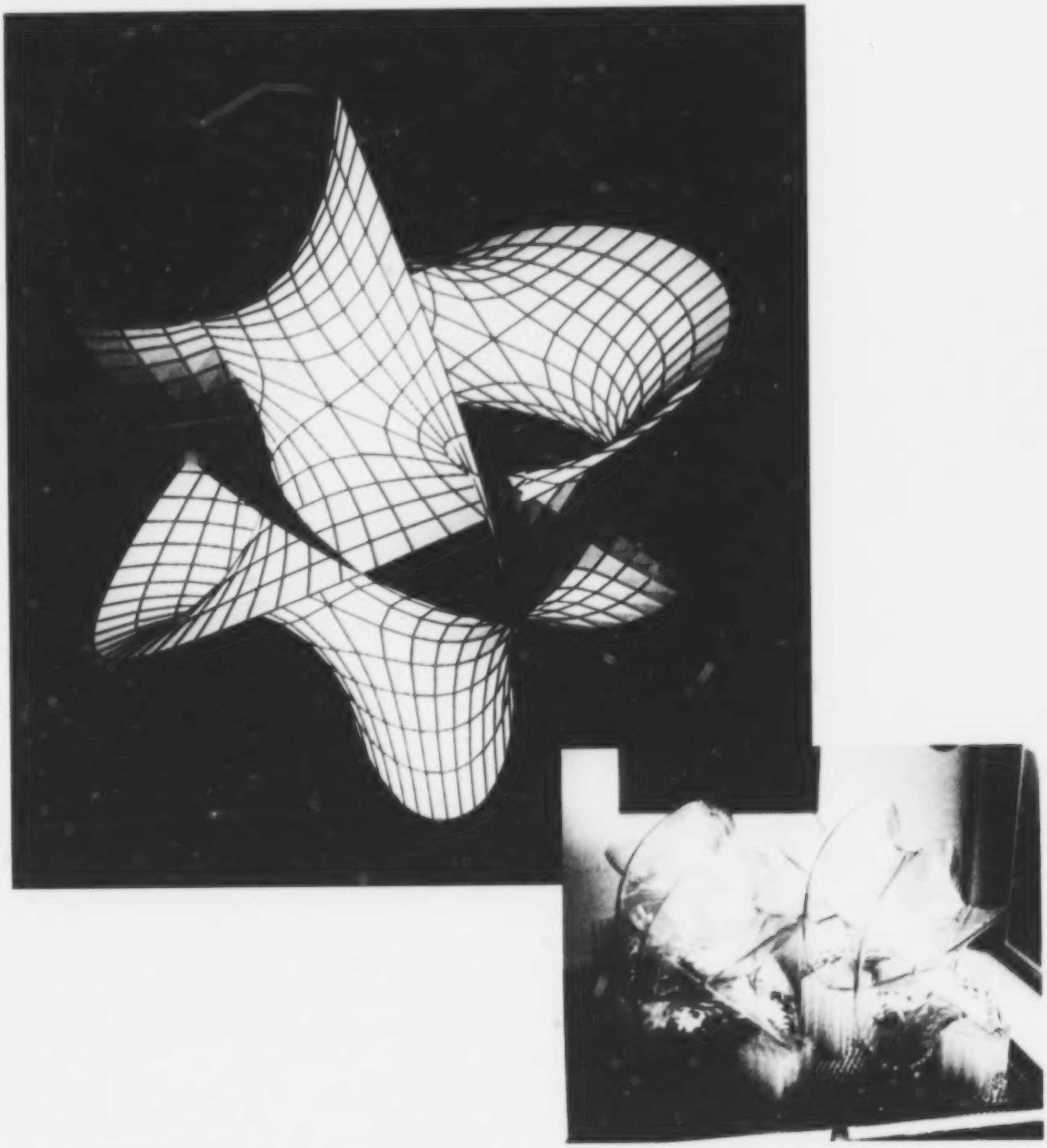


Figure 1. *A 3-dimensional projection of the complex Fermat surface* $u^3 + v^3 = 1$, *rendered with computer graphics (top) and through stereolithography as a plastic sculpture (bottom). (Figure courtesy of Stewart Dickson.)*

"A Truly Remarkable Proof"

A torrent of electronic mail poured from Cambridge, England, on the morning of June 23, 1993. Mathematicians at a conference on number theory at the Isaac Newton Institute, a mathematical research center at the University of Cambridge, raced to tell their colleagues around the world some stunning news: Andrew Wiles, a number theorist at Princeton University, had just finished presenting a proof of Fermat's Last Theorem.

Wiles, it seemed, had solved mathematics' most famous open problem. Fermat's Last Theorem is a deceptively simple statement: The equation $x^n + y^n = z^n$ has no solutions in positive integers x, y, and z if the exponent n is greater than 2. The theorem was jotted down by the French mathematician Pierre de Fermat around 1637, in the margin of a math book, along with a tantalizing comment: "I have discovered a truly remarkable proof, which this margin is too small to contain."

Countless mathematicians over the last 350 years have tried and failed to supply the missing proof. Prize money has even been offered for a solution. Curiously, by the usual standards of mathematics, the theorem itself is of little consequence: Unlike other famous unsolved problems in mathematics, Fermat's Last Theorem has no important corollaries. Rather, the problem's significance stems mainly from the theoretical machinery researchers have developed in trying

Andrew Wiles. (Photo courtesy of Denise Applewhite and Princeton University.)

BEST COPY AVAILABLE

Number theorists would like to know whether *all* elliptic curves are modular. The Taniyama–Shimura conjecture says they are.

to solve it. Indeed, most mathematicians long ago gave up working directly on Fermat's Last Theorem itself. Then Wiles dropped his bombshell in Cambridge.

The news lit up the mathematical world. It also grabbed the media's attention, as mathematical stories seldom do. Wiles's proof made the front page of the *New York Times*. It made *Time* and *Newsweek*. It made the NBC Nightly News ("Be still, my heart," said NBC's Tom Brokaw).

Experts who attended Wiles's lectures at the Newton Institute expressed confidence in the strategy of his proof, and amazement at the mathematical *tour de force* it represented. Still, mathematicians accept no proof as correct until it's been thoroughly checked—especially when the problem has the stature of Fermat's Last Theorem. And after the initial celebration had subsided and experts began meticulously poring over Wiles's 200-page manuscript, problems with the proof appeared. Most were minor, but one was not.

In early December, Wiles posted an e-mail message acknowledging a gap in the reasoning near the end of his proof. As this volume of *What's Happening* goes to press, the gap remains. Fermat's Last Theorem is still an open problem.

Yet number theorists continue to praise Wiles's work. "When people finally see this manuscript, they're just going to be bowled over completely," says an admiring Ken Ribet of the University of California at Berkeley. That's because Wiles's work, while aiming to prove Fermat's Last Theorem, advances number theory across a broad front. Indeed, the main focus of his work is not Fermat's Last Theorem itself, but one of the central problems in modern number theory, an assertion known as the Taniyama–Shimura conjecture.

To explain the Taniyama–Shimura conjecture and its relation to Fermat's Last Theorem requires a brief digression on the subject of elliptic curves. Roughly speaking, an elliptic curve is the set of solutions to a cubic equation in two variables. A typical equation, such as $y^2 = x(x - 3)(x + 32)$, sets the square of one variable equal to a cubic expression in the other. Number theorists are particularly interested in "rational points" on elliptic curves: solutions in which both x and y are rational numbers (see Figure 2).

One way to study the rational points on an elliptic curve is to look at the curve not in the ordinary system of real numbers, but in an infinite collection of *finite* number systems. In each finite system, the elliptic curve's cubic equation can be solved explicitly, and the number of solutions tallied. Number-theoretic properties of the original elliptic curve are reflected in solutions of the cubic equation in these finite systems.

Things work best when the elliptic curve in question is "modular." Modularity is a complicated, technical condition, but essentially it means that there is a formula for the number of solutions of the curve's cubic equation in each finite number system. Many elliptic curves are known to be modular, and the condition can be checked computationally for individual curves. Number theorists would like to know whether *all* elliptic curves are modular. The Taniyama–Shimura conjecture says they are.

First formulated in 1955 by the Japanese mathematician Yutaka Taniyama, and later refined by Goro Shimura at Princeton University, the Taniyama–Shimura conjecture was—and still is—a bold and striking characterization of elliptic curves. In its full technical glory, the conjecture asserts that every elliptic curve is associated with a particular kind of function known as a modular form: this links two seemingly unrelated branches of number theory. The idea that there's a bridge between elliptic curves and modular forms "really pervades lots of things that we do" in modern number theory, says Ribet. And unlike Fermat's Last Theorem, the Taniyama–Shimura conjecture has a host of immediate consequences.

Fermat's Last Theorem is one of them.

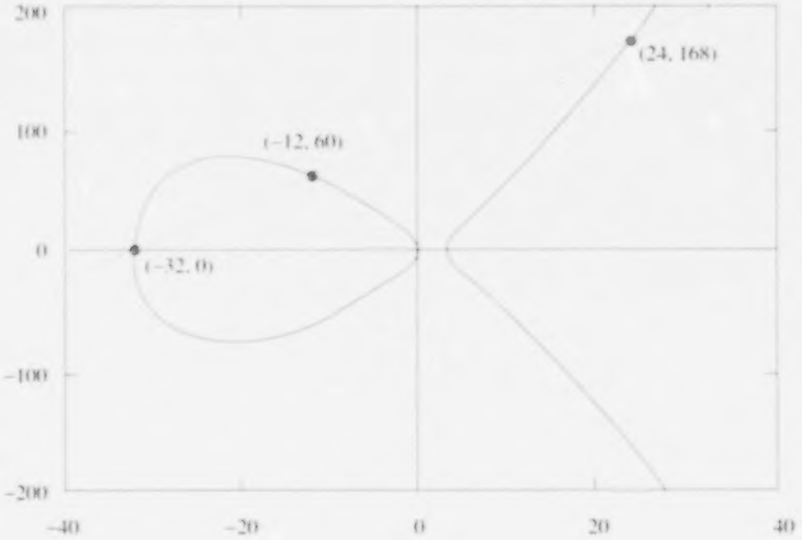


Figure 2. *The elliptic curve* $y^2 = x(x-3)(x+32)$ *has many rational points. A line connecting any two of them intersects at a third.*

The connection between Fermat's "simple" problem and the theory of elliptic curves came as a surprise when, in 1985, Gerhard Frey at the University of the Saarland in Saarbrücken, Germany, had the idea that any counterexample to Fermat's Last Theorem could be used to construct a counterexample to the Taniyama–Shimura conjecture. Specifically, Frey proposed, if $a^n + b^n = c^n$ for positive integers a, b, and c and an exponent n greater than 2, then the elliptic curve with cubic equation $y^2 = x(x - a^n)(x + b^n)$ cannot be modular.

Frey's idea hinged on a technical result, which Jean-Pierre Serre at the Collège de France in Paris formulated as a precise conjecture. A year later, Ribet proved Serre's conjecture. This established Fermat's Last Theorem as a consequence of the Taniyama–Shimura conjecture.

Ribet's result gave mathematicians a brand new way of thinking about Fermat's Last Theorem and a new reason to work on the Taniyama–Shimura conjecture. Actually, it's not necessary to establish the Taniyama–Shimura conjecture in full generality in order to deduce Fermat's Last Theorem; it's enough to prove it for a class known as semistable curves. This was the starting point for Wiles's attack on the problem.

Wiles, who was already well known as an expert in the theory of elliptic curves, went to work full time on the Taniyama–Shimura conjecture. To avoid undue publicity he kept only one colleague at Princeton, Nicholas Katz, abreast of developments. Finally, in June, he asked to give three talks at the Newton Institute number theory conference. John Coates of Cambridge University, who was Wiles's

Ribet's result gave mathematicians a brand new way of thinking about Fermat's Last Theorem and a new reason to work on the Taniyama–Shimura conjecture.

In his third lecture, Wiles announced his major result: The Taniyama–Shimura conjecture is true for semistable elliptic curves.

thesis advisor at Cambridge in the mid-1970s, scheduled him to speak on Monday, Tuesday, and Wednesday, June 21–23, 1993.

The audience could tell just from the title of his lectures — "Elliptic curves, modular forms, and Galois representations" — that Wiles had important news to impart, perhaps pertaining to Fermat's Last Theorem. (All three items in Wiles's title are key ingredients in Ribet's 1986 result). What Wiles began laying out, says Ribet, was "a complete revelation which is really still shaking number theory": a new method for proving that elliptic curves are modular.

Wiles's theory builds on results of many other mathematicians, including recent work by Matthias Flach at the University of Heidelberg, Victor Kolyvagin at the Steklov Institute in Moscow, Barry Mazur at Harvard University, Ribet, and Karl Rubin at the Ohio State University. The new method for proving modularity is extremely powerful. In essence, it reduces the problem of showing that the Taniyama–Shimura conjecture holds for particular elliptic curves to the proof of a single algebraic inequality. That by itself is a "fantastic new result," says Rubin. For a large class of elliptic curves, the inequality is easy to verify. In his first two lectures, Wiles outlined how the new method proves the Taniyama–Shimura conjecture for one infinite family of elliptic curves, another enormous advance in its own right. His lectures left the audience wondering if he had left the family of semistable curves — those that pertain to Fermat's Last Theorem — for last.

Left to right: John Coates, Andrew Wiles, Ken Ribet, and Karl Rubin at the Isaac Newton Institute in Cambridge, England, after Wiles's historic talk. (Photo courtesy of Ken Ribet.)

He had. In his third lecture, Wiles announced his major result: The Taniyama–Shimura conjecture is true for semistable elliptic curves. Almost as an afterthought, he noted the long-awaited corollary: Fermat's Last Theorem. It took a moment for the announcement to sink in. Then the audience burst into applause.

"The logic of his argument is utterly compelling," Ribet said at the time. Other number theorists agreed that Wiles had cleared many of the technical hurdles on the way to a proof of the Taniyama–Shimura conjecture and had set a new agenda for the theory of elliptic curves. However, the review process has revealed a gap near the end of the proof: The calculations that verify the crucial inequality, which are easy in some cases, turn out to be not so easy for the class of semistable curves.

Experts believe that Wiles's basic strategy for the calculations is sound, even if the details don't yet fit together.

It's not unusual for a long, complicated mathematical proof to contain an error, Wiles's colleagues are quick to point out (it's not even unusual for a *short* proof to be mistaken). Nobody knows how long it will take to fill the gap. Still, says Rubin, "it's hard to believe that the proof of Fermat's Last Theorem is not closer."

Nobody knows how long it will take to fill the gap.

Fermat's Last Theorem is True (for Exponents up to 4,000,000)

Fermat's Last Theorem states that the equation $x^n + y^n = z^n$ has no solutions in positive integers x, y, and z if the exponent n is greater than 2. Could the theorem be true for some exponents and false for others? Mathematicians have made much progress in the last 350 years in showing that if any counterexamples exist, the numbers involved are colossal.

Although he never found room in the margin or anywhere else for a general proof, Fermat did write down a proof of his famous theorem for the special case $n = 4$. Over a hundred years later, the Swiss mathematician Leonhard Euler dispatched the case $n = 3$. In the 1820s and 1830s, the theorem was proved for exponents 5 and 7. (It's enough to prove Fermat's Last Theorem for *prime* exponents. For example, if $x = a$, $y = b$, and $z = c$ solve the equation $x^6 + y^6 = z^6$, then $x = a^2$, $y = b^2$, and $z = c^2$ solve $x^3 + y^3 = z^3$.)

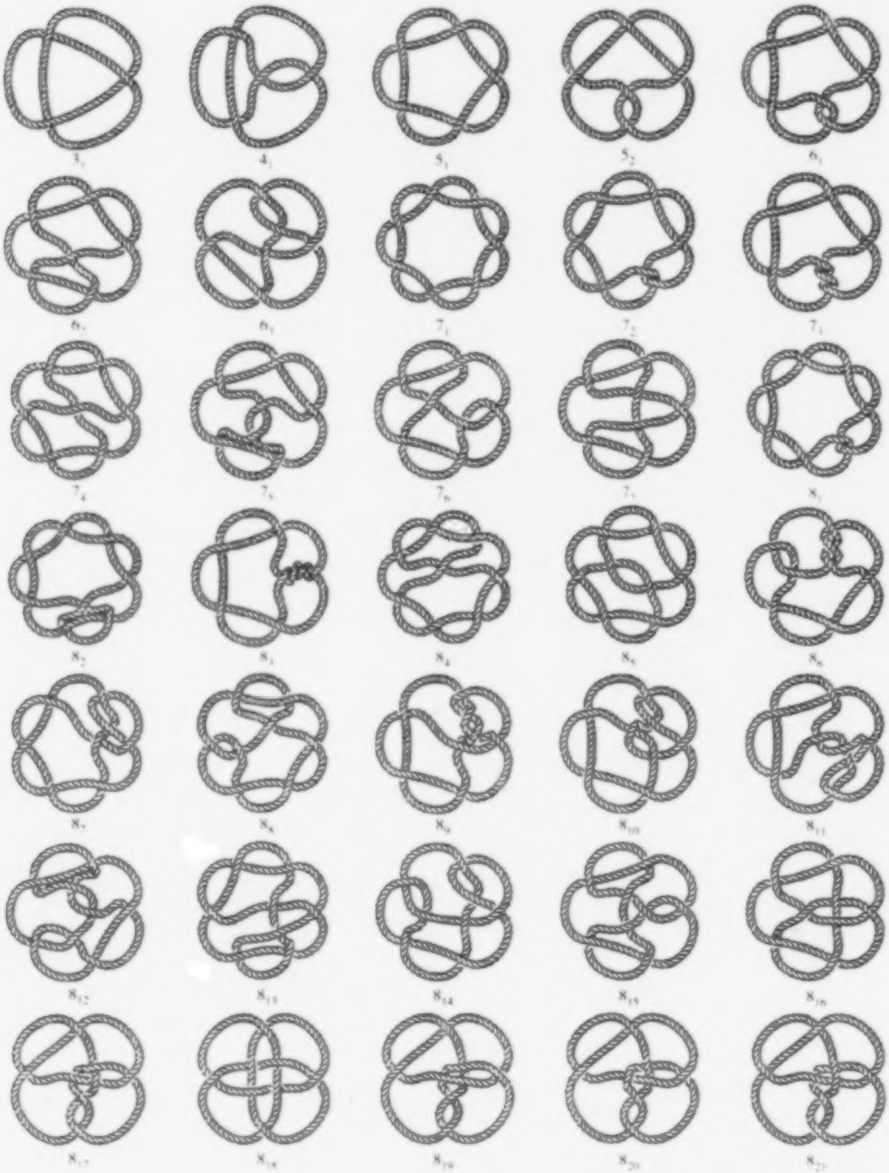
In the 1840s, the theory took a giant leap forward. By introducing some potent new ideas, Ernst Eduard Kummer was able to prove Fermat's Last Theorem for all prime exponents up to 100, with the exception of three "irregular" primes. In Kummer's theory, primes are classified as either regular or irregular. Fermat's Last Theorem, the theory says, is true for all regular primes. Regular primes are believed to be more common than irregular ones, constituting roughly 60% of all primes. Ironically, though, while it's known that there are infinitely many irregular primes, the same statement (while undoubtedly true) has never been proved for regular primes.

Later improvements in Kummer's theory made it possible to handle irregular primes separately, on a case-by-case basis. In effect, the theory reduces the proof of Fermat's Last Theorem for individual exponents to a straightforward, though lengthy, computation—tailor-made for modern computers.

In the 1970s, Sam Wagstaff at Purdue University used this approach to establish Fermat's Last Theorem for all exponents up to 125,000. Recently, four researchers have pushed the computational approach into the millions. Using refinements of Kummer's basic theory to speed up the calculation, Joe Buhler at Reed College in Portland, Oregon, and Richard Crandall at NeXT Computer Inc., in Redwood City, California, with help from Tauno Metsänkylä and Reijo Ernvall at the University of Turku in Finland, have verified Fermat's Last Theorem for all exponents up to 4 million. Their results, which appeared in 1993 in the journal *Mathematics of Computation*, also support the conjecture regarding the ratio of regular to irregular primes: Out of 283,145 primes up to 4 million, 171,548, or 60.59%, are regular.

Extending Fermat's Last Theorem beyond the 4 million mark is certainly possible, says Buhler, but doing so will require developing new computational techniques. If Wiles succeeds in filling the gap in his proof, that won't be necessary.

Knots to 8 Crossings


3_1
4_1
5_1
5_2
6_1
6_2
6_3
7_1
7_2
7_3
7_4
7_5
7_6
7_7
8_1
8_2
8_3
8_4
8_5
8_6
8_7
8_8
8_9
8_{10}
8_{11}
8_{12}
8_{13}
8_{14}
8_{15}
8_{16}
8_{17}
8_{18}
8_{19}
8_{20}
8_{21}

David Broman and Charlie Gunn

Figure 1. *Reprinted with permission from* Supplement to Not Knot *by David Epstein and Charlie Gunn, published by A K Peters, Ltd.*

From Knot To Unknot

Alexander the Great didn't mess around. As legend has it, the Macedonian king decided to try his luck with the fabled Gordian knot, a tough length of cornel bark wrapped tightly around the pole of an ox cart. It was said that the person who succeeded in untying this knot was destined to rule the world (meaning, at the time, Persia). A man of action rather than dexterity or patience, Big Al unsheathed his sword — and the rest, as they say, is history.

Modern mathematicians are also drawn to the problem of undoing knots, although their motives — and their techniques — are quite different from Alexander's. In the last few years, researchers using two different approaches have come to understand better just what it takes to unknot a knot.

A mathematical knot is basically just a closed curve that winds through 3-dimensional space, like an electrical extension cord that's been tangled up and then plugged into itself. The theory of these meandering curves has taken off in the last decade. "Knot theory for a long time was a little backwater of topology," notes Joan Birman, an expert in the subject at Columbia University. "It's now been recognized as a very deep phenomenon in many areas of mathematics." And it's not just mathematics where knot theory is playing a larger role: molecular biologists, for example, are using it to help untangle some of the geometric secrets of DNA (see box page 13).

One key problem in knot theory is to decide whether one knot can be deformed into another — in particular, to tell whether a given knot really isn't knotted at all. That may sound like a straightforward, even trivial, problem. But it's really as difficult to deal with as a snarled-up fishing line. The main difficulty is that there are infinitely many ways to deform any knot, and they all must be ruled out in order to show that two knots are indeed different.

Because it's hard to draw truly 3-dimensional pictures, knots are commonly represented by projections onto a plane. Such a picture, called a "knot diagram," can be thought of as tracing the path of a tangled extension cord that's been dropped onto the floor: places where the curve is broken are called crossings. The number of crossings depends on how the cord has been dropped, and can be quite large — but every knot has a diagram with a minimal number of crossings. Knot theorists have constructed elaborate tables of knots arranged according to this number (see Figure 1). These tables were begun in the 1890s by the British mathematician P. G. Tait, who was inspired by Lord Kelvin's theory that atoms were "knotted vortices" in the ether. (Kelvin's idea did not survive, but surprisingly, knot theory has re-emerged in physics, this time in an area known as quantum field theory.)

In the 1920s, the German mathematician Kurt Reidemeister showed that any deformation of a knot can be achieved by a sequence consisting of three types of moves (see Figure 2). This gives a combinatorial flavor to the topological problem of classifying knots, but it does not automatically solve the problem, because there are no set rules that specify the order in which the moves should be applied. For example, you might think that if a knot can be deformed into the "unknot" (the knot theorist's word for the circle), the deformation could be done without ever increasing the number of crossings. That's not true: For some diagrams, the crossing number must go up before it can come down (see Figure 3).

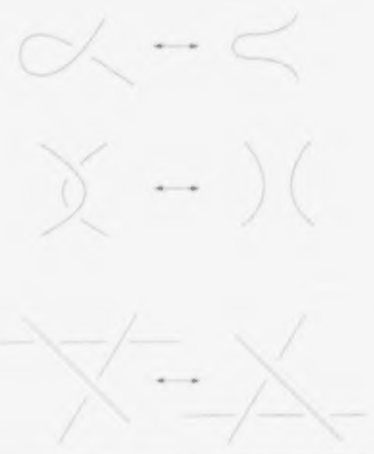


Figure 2. *The three types of Reidemeister moves.*

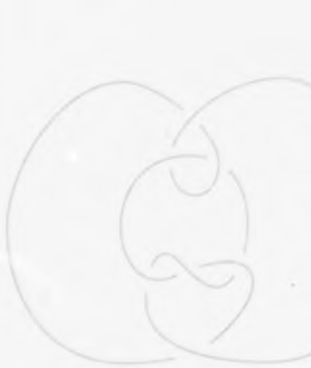


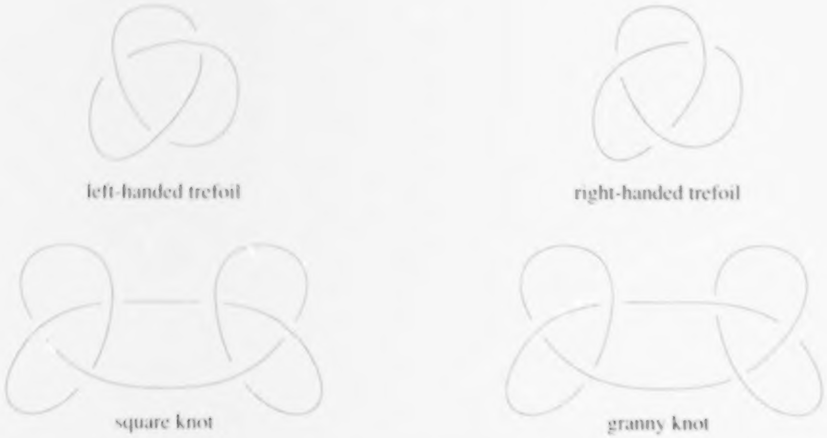
Figure 3. *A "nasty" unknot that can only be unknotted by first increasing the number of crossings.*

In the last few years, researchers using two different approaches have come to understand better just what it takes to unknot a knot.

So how *do* mathematicians decide whether two knots are different? Knot theorists' favorite approach has been to compute "invariants": numerical or algebraic expressions assigned to a knot that don't change when the knot is deformed. One of the earliest invariants, also dating back to the 1920s, is known as the Alexander polynomial (named after the American mathematician John Alexander, not Alexander the Great). Though defined topologically, the Alexander polynomial can be derived from the pattern of under- and over-crossings in a knot diagram. Most important, if two knots have different Alexander polynomials, then they are necessarily different knots. For example, the trefoil knot, whose polynomial is $x^2 - x + 1$, differs from the square knot, whose polynomial is $(x^2 - x + 1)^2$—and both differ from the unknot, whose polynomial is the constant 1 (see Figure 4). However, different knots need not have different Alexander polynomials. The granny knot, for example, has the same polynomial as the square knot. Likewise, the right- and left-handed trefoil knots share the same polynomial even though it's impossible to deform one into the other.

For a long time, though, the Alexander polynomial was one of the few tools topologists had for telling knots apart. Then in 1984, Vaughan Jones, a mathematician at the University of California at Berkeley, discovered a new polynomial invariant. Jones's polynomial turned out to be more powerful than Alexander's at distinguishing different knots. It also revealed startling new connections between knot theory and mathematical physics. More recently, Viktor Vassiliev at the Independent University of Moscow has introduced a whole new class of invariants based not on the topology of individual knots but on the structure of the space of *all* closed curves, even those that pass through themselves (such curves are viewed as degenerate, or "singular," knots). Birman and Xiao-Song Lin at Columbia University have found deep connections between Vassiliev's invariants and the Jones polynomial.

The Alexander and Jones polynomials are easy to compute, but they don't refer to anything that can be seen geometrically in a knot diagram. The minimal crossing number for a knot, on the other hand, refers explicitly to something that



left-handed trefoil

right-handed trefoil

square knot

granny knot

Figure 4. *The square knot is formed by joining a right- and left-handed trefoil; the granny knot is formed by joining two right-handed (or two left-handed) trefoils.*

can be seen. So does the "unknotting number," which is the least number of times you need to "cheat" by passing a knot through itself in order to untie it. But these invariants can be hard to compute.

Theorists generally compute a knot's minimal crossing number by a process of elimination: First they find a diagram that seems to have the fewest crossings; then they show that the knot is different from every knot with fewer crossings, typically by comparing Alexander or Jones polynomials. The second step, however, requires a complete list of knots with smaller crossing numbers. So far that list is complete only up to crossing number 14.

The unknotting number is even harder to compute. You might think you could compute it by taking any diagram and finding the smallest combination of crossings which, if the knot is passed through itself at those points (so that underpasses become overpasses and vice versa), will untie the knot. Unfortunately, that doesn't necessarily give the right answer—it only puts an upper bound on the unknotting number. For example, Figure 5 (top) shows a knot diagram—in fact, one with a minimal number of crossings—which cannot be untied by changing fewer than three crossings. But Figure 5 (bottom) shows the same knot in a diagram that can now be untied with just two cheats. In other words, to find the fewest crossings to change, it may be necessary to take a simple looking diagram and redraw it to look more complicated; and there seems to be no bound on how much more complicated a knot diagram may need to look before it exhibits the correct unknotting number.

Until the recent breakthrough, theorists had no general method for computing unknotting numbers, except when the value happens to be 1 (if a knot can be untied with a single cheat, then its unknotting number must be either 1 or 0, so one need only check whether the knot was already unknotted). But researchers have recently proved results that allow knot theorists to compute unknotting numbers exactly for many more knots, and obtain useful lower bounds on unknotting numbers for *all* knots.

Working on problems in 4-dimensional topology, Peter Kronheimer at Oxford University and Tomasz Mrowka at the California Institute of Technology have proved a 40-year-old conjecture, due to John Milnor, about unknotting numbers for a special class of knots. Milnor's conjecture specifies the unknotting number for all "torus" knots. These knots come from curves that are drawn on a torus, which is what mathematicians call a donut. To tie a (p, q)-torus knot, wrap a string p times through the hole in a donut, stretching the string so it goes q times around the donut itself before you tie the ends of the string together; then eat the donut (see Figure 6).

Milnor, who was also mainly interested in 4-dimensional topology, conjectured that the unknotting number for such a knot is always $(p-1)(q-1)/2$ (p and q can't both be even; in fact, in order for the knot to be drawn on the torus without intersecting itself, p and q can't have any common divisor greater than 1). For example, Milnor's conjecture says that the (101, 3)-torus knot has unknotting number 100. Given the difficulty knot theorists have had computing the unknotting number when its value is greater than 1, Milnor's conjecture seems almost miraculous.

How, you might wonder, does 4-dimensional topology get mixed into the theory of knots? The trick is to view the deformation of a knot as occuring in time, which adds a fourth dimension to space. "If we think of a space-time picture of what is going on, the moving curve in space sweeps out a 2-dimensional surface in space-time," Kronheimer explains. Topologically, the 2-dimensional surface is

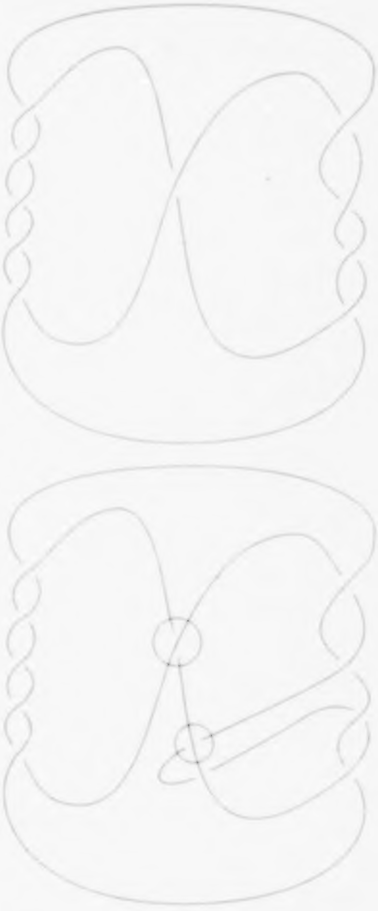


Figure 5. *The diagram at top cannot be unknotted with fewer than three changes of crossings, but the modified diagram below can be unknotted with only two (indicated by circles).*

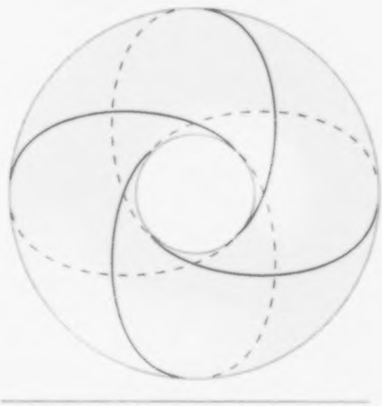


Figure 6. *A (4,3)-torus knot.*

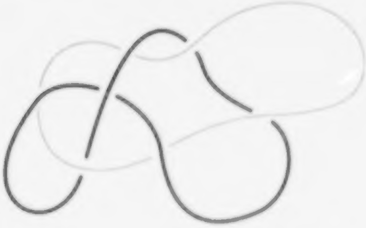


Figure 7. *Two unknots that are "linked" together.*

like a cylinder, he adds, but "because the original curve is knotted, the cylinder sits in space–time in a rather complicated way."

If the deformation includes a cheat, then the space–time surface intersects itself at that point. Kronheimer and Mrowka were studying the general theory of self-intersecting, or "immersed," surfaces in 4-dimensional space. Their research is based on far-reaching ideas introduced in the early 1980s by Simon Donaldson at Oxford University, who borrowed techniques from theoretical physics to analyze the structure of 4-dimensional spaces. The result for knots came as a kind of 3-d bonus. "We weren't really aiming at Milnor's [conjecture]," Kronheimer says.

More recently, Lee Rudolph at Clark University in Worcester, Massachusetts, has shown that Kronheimer and Mrowka's results also prove a generalization of Milnor's conjecture due to Daniel Bennequin at the University of Strasbourg in France, which provides a lower bound on the unknotting number for all knots, and all "links" as well. (A link is simply a set of knots that are tangled together, as in Figure 7.) Bennequin's conjecture requires drawing the knot (or link) diagram in a particular configuration known as a braid and distinguishes between "positive" and "negative" crossings (see Figure 8). If a braid with M strands and R components (which is 1 for a knot, and greater than 1 for a link) has P positive and N negative crossings, then Bennequin's conjecture asserts that the unknotting number U satisfies the inequalities $|P - N| \leq 2U + M - R \leq P + N$. If all crossings have the same sign (say positive), then Bennequin's conjecture gives an exact value for the unknotting number. In particular, it turns out that a (p, q)-torus knot can be drawn as a braid with p strands and $(p - 1)q$ positive crossings, so Milnor's formula falls out of Bennequin's conjecture. (Actually, only the lower bound in Bennequin's conjecture required proof: the upper bound was proved by Michel Boileau at the University of Toulouse and Claude Weber at the University of Geneva in 1983, shortly after the conjecture appeared.)

As if one proof weren't enough, William Menasco, a knot theorist at the State University of New York at Buffalo, has also proved Bennequin's conjecture, using completely different methods. (This parallels the legend of Alexander. According to some accounts, the Great one didn't draw his sword, but instead removed the pole on which the Gordian knot was tied, leaving the knot to fall apart of its own accord.) Working independently at about the same time as Kronheimer and Mrowka, Menasco actually proved a stronger version of Bennequin's conjecture, one that looks more closely at the distinction between positive and negative crossings in a knot or link.

In Menasco's theorem, the unknotting number is replaced with positive and negative variants. The positive unknotting number, U_+, is defined as the minimal number of positive crossings that must be changed to negative ones in order to untie a knot, regardless of how many negative crossings must be changed to positive. (The negative unknotting number, U_-, is defined similarly.) Menasco showed that U_+ satisfies the inequality $P - N \leq 2U_+ + M - R$, provided that $P \geq N$. (A similar inequality holds for U_- if $P \leq N$). Since the original unknotting number U is never less than U_+ or U_-, Menasco's inequalities together imply Bennequin's conjecture.

Menasco's proof is strictly 3-dimensional. Like Kronheimer and Mrowka's proof, it is based on a careful study of immersed surfaces, but in this case the surfaces are deformed disks bounded by knots, all situated in ordinary 3-dimensional space. The proof is "very geometric" and involves "a lot of picture drawing," Menasco says, adding that his approach uses "low-tech mathematics" compared

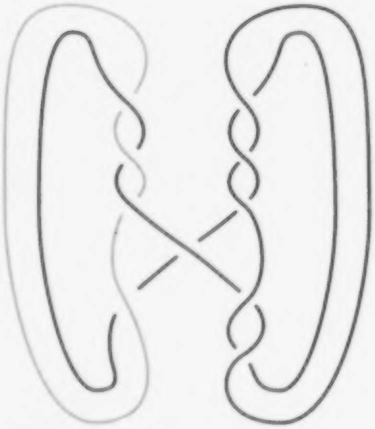


Figure 8. *A 4-strand braid with 2 knot components, 8 positive crossings, and 2 negative crossings. (In a positive crossing, the overpass goes from upper left to lower right; in a negative crossing, it goes from upper right to lower left.)*

to the methods employed by Kronheimer and Mrowka. Birman, who has collaborated with Menasco on an extensive study of links and braids, disagrees. "The proof that he found is very hard," she says. "Some of the things that he did are extremely difficult to visualize. His ability to visualize 3-dimensional geometry is rather extraordinary."

Birman is enthusiastic about the new results on the unknotting number. "It's the beginning of a real theory of this mysterious number," she notes. She is also pleased that there are two proofs. "It's quite wonderful that two such widely different techniques could lead to the same result," she says. "I think it's evidence of the unity of mathematics."

But that's what knots are good for: Tying things together.

According to some accounts, Alexander the Great didn't draw his sword, but instead removed the pole on which the Gordian knot was tied, leaving the knot to fall apart of its own accord.

The Knotted Helix

Mathematicians aren't the only ones excited by the latest results in knot theory. Molecular biologists, too, are eager to get in on the action.

"For me it's great," says Sylvia Spengler, a molecular biologist at the University of California at Berkeley. "It gives me insight on how frequently an enzyme had to act."

Spengler is one of a growing group of researchers applying theorems from topology to the chemistry of life. That may seem like a stretch—but stretching is what topology is all about. Biologists have long known that DNA is not only wound in a double helix, but also tightly coiled inside the nucleus of the cell. But only recently have researchers begun to understand the details of what they call supercoiling.

Supercoiling is found, for example, in circular DNA, a form of the macromolecule that occurs in bacteria and yeast. The flexible molecule need not look like a geometric circle, though; it may even be knotted. Knotting—and unknotting—is caused by enzymes called topoisomerases. These enzymes cut the strand of DNA at one point, pass another part of the strand through the gap, and then reseal the cut—exactly what's called for in the unknotting number theorem.

De Witt Sumners, a knot theorist at Florida State University who collaborates with Spengler and others on topological aspects of molecular biology, points out that the unknotting number is a lower bound for the number of times the topoisomerase has to act. "If you have really complicated products that have a large unknotting number, it's going to take the enzyme a while to produce those," he explains.

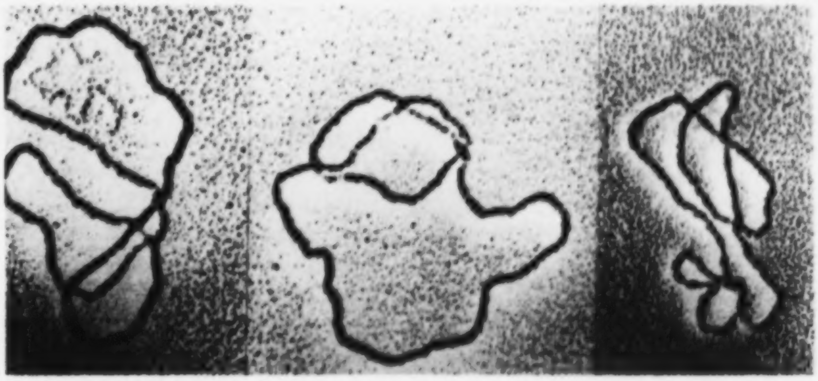


Figure 9. *Three strands of knotted circular DNA. (Photo courtesy of Sylvia Spengler, University of California, Berkeley, and Frank Dean, Rockefeller University.)*

BEST COPY AVAILABLE

New Wave Mathematics

Philip Rosenau

John Scott Russell knew he was on to something one August day in 1834, when he chased a peculiar "heap of water" down the Edinburgh Glasgow canal. When a boat being pulled by horses suddenly stopped, the wave "rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth, and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed." The wave, he noticed, was approximately 30 feet long and a foot and a half high, and traveled at 8 or 9 miles per hour. Russell followed it on horseback for a couple of miles until it finally disappeared. "Such," he later wrote, "was my first chance interview with the singular and beautiful phenomenon which I have called the Wave of Translation."

At the time, Russell's observation was considered an anomaly: it was even greeted with disbelief. These days, the theory of solitary waves is a well developed subject, with close ties to mathematical physics. But even so, there are still surprises and potential applications waiting in the wings. One surprise surfaced recently when Philip Rosenau, a theorist at the Technion in Haifa, Israel, discovered a class of waves so solitary that two of them can move along within a hair's breadth of each other yet remain blissfully unaware of each other's existence. Rosenau and Mac Hyman, a mathematician in the Theoretical Division at Los Alamos National Laboratory, have been chasing and observing these compact waves not on horseback, but by means of high-speed computation.

Although their findings are, so far, strictly mathematical, applications of Rosenau and Hyman's compact waves may not be far off. For years, solitary waves have been considered as promising carriers of digital information on optical fibers, because they can, in principle, travel forever without losing their shape. Compact waves' ability to travel close together without interfering with each other might offer even further advantages.

Mathematicians in the nineteenth century were slow to come to grips with Russell's solitary wave, in part because the prevailing theory of wave motion was locked into a particular partial differential equation called the "wave equation," which is still used to describe all kinds of undulatory phenomena, from water waves to sound waves to quantum-mechanical waves (the last, of course, being a twentieth-century innovation). According to the wave equation, Russell's "heap of water" couldn't sustain itself: It would immediately begin to break apart, as high-frequency components raced out in front, leaving lower-frequency components further and further behind—exactly what Russell didn't see.

The wave equation also failed to explain another observation Russell made, this time when he re-created solitary waves in his laboratory by dropping weights into a long rectangular tank of water. The taller the wave, Russell found, the faster it moved. For explaining this behavior, the wave equation is no help at all: The wave equation is *linear*, and for phenomena described by linear equations, the height of things does not affect how they change in time.

By the end of the nineteenth century, however, an adequate theory for solitary waves had been developed, in the form of a modified wave equation known as the Korteweg de Vries, or KdV, equation. Derived from basic equations of fluid dynamics, the KdV equation describes how waves propagate down a channel with rectangular cross section. It differs from the ordinary wave equation in one critical respect: the KdV equation has a nonlinear term.

Rosenau and Hyman's equations go even further. The two theorists tinkered with the nonlinear term in the KdV equation; more important, they added a second nonlinearity, this time in a part of the equation known as the dispersion term (see box). The inspiration for making the dispersion term nonlinear came from Rosenau's studies of liquid drops, such as raindrops running down a window pane.

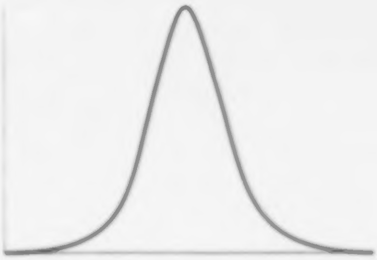


Figure 1. *A traveling-wave solution of the KdV equation.*

A Tale of Three Equations

The wave equation, the KdV equation, and the compacton equation are all roughly similar in form, but the differences are critical. All three equations can be thought of as describing the up-and-down motion of a string of corks floating in a narrow channel of water, such as a long, thin trough. Mathematically, the trough is infinitely long and infinitely thin, so that each cork can be identified by a single variable, say x, which specifies its location along the length of the trough. The corks' up-and-down motion is described by a function of two variables: $u(x,t)$ is the height of the cork at point x and time t.

The traditional, linear wave equation has the form $u_t + u_x + u_{xxx} = 0$. The first term, u_t, is the derivative of u with respect to t—that is, the speed at which a cork is going up or down. The middle term, u_x, is the derivative of u with respect to x: it describes the slope of the wave at each point—that is, how much higher or lower each cork is than its neighbors at a particular moment. The final term, u_{xxx}, is the third derivative of u with respect to x. This is the "dispersion" term: because of it, traveling-wave solutions with different wavelengths propagate at different speeds. These solutions have the form $u(x,t) = \sin((x - ct)/\ell)$ with $c = (\ell^2 - 1)/\ell^2$. The parameter ℓ is the wavelength, while c is the speed at which the wave propagates (i.e., the speed at which a particular crest of the wave moves). Because the equation $u_t + u_x + u_{xxx} = 0$ is linear, a general solution can be formed by adding these traveling-wave solutions together. But any such combination will produce a wave that changes shape over time, because the different components move at different speeds.

The KdV equation has the form $u_t + (u^2)_x + u_{xxx} = 0$. Changing the middle term—squaring the u before taking its first derivative—has a profound effect. Simple sine waves are no longer solutions: instead, the traveling-wave solutions have the form $u(x,t) = (3c/2)\text{sech}^2((x - ct)\sqrt{c}/2)$. The shape of the solution is a single, smooth hump (see Figure 1). What's more, both the height and the "width" of a wave are completely determined by the speed c at which it travels: Taller waves move faster than shorter ones, quite unlike waves governed by linear equations. Whatever shape a wave governed by the KdV equation has initially, it will eventually—and usually quite quickly—break up into a train of these basic shapes, with the tallest waves out front.

Rosenau and Hyman's compacton equations make the dispersion term in the KdV equation nonlinear as well. The simplest compacton equation has the form $u_t + (u^2)_x + (u^2)_{xxx} = 0$. (More generally, Rosenau and Hyman have studied equations of the form $u_t + (u^m)_x + (u^n)_{xxx} = 0$.) This time, the traveling-wave solutions have the form $u(x,t) = (4c/3)\cos^2((x - ct)/4)$ for $-2\pi < x - ct < 2\pi$ and $u(x,t) = 0$ if $|x - ct| \geq 2\pi$. As with the KdV equation, the height of a compacton depends on its speed, but compactons have the same width, namely 4π. Waves in any initial shape also decompose into a train of compactons. However, numerical evidence suggests that when compactons separate, as they do after a collision, they leave behind them an apparently infinite wake of tiny ripples. Rosenau and Hyman are still trying to fathom the nature of these ripples.

Most nonlinear equations cannot be solved exactly—that's one of the advantages linear equations hold. But some can. The KdV equation and the new nonlinear dispersion equations turn out to be among them. The basic traveling-wave solution of the KdV equation involves a special function known as a hyperbolic secant. (The

What happens when a fast-moving wave overtakes a slow-moving wave?

hyperbolic functions, often seen in introductory calculus classes, are closely related to the trigonometric functions commonly studied in high school.) Rosenau and Hyman's equations, by contrast, have various traveling-wave solutions, ranging from parabolic arcs to cosine waves. But all of these waves, whether generated by the KdV equation or by Rosenau and Hyman's equations, share one particularly striking feature: The speed at which a solitary wave moves is proportional to its height—just as Russell had seen in his laboratory.

That feature raises an interesting question: What happens when a fast-moving wave overtakes a slow-moving wave? In the 1960s, Martin Kruskal at Princeton University and Norman Zabusky at Bell Telephone Laboratories in Whippany, New Jersey (both now at Rutgers University in New Brunswick, New Jersey), found a surprising answer. When a tall solitary wave overtakes a shorter one, the two do not merely merge. Nor do they break each other apart. Instead, after a brief but passionate encounter, the two waves separate, each with the same size and shape it had before. The only evidence they ever met is a "phase shift": The taller wave is pushed slightly ahead of where it would otherwise have been, while the shorter wave is held slightly back. Because the solitary waves retain their separate identities, much as colliding particles do, Kruskal and Zabusky dubbed them "solitons."

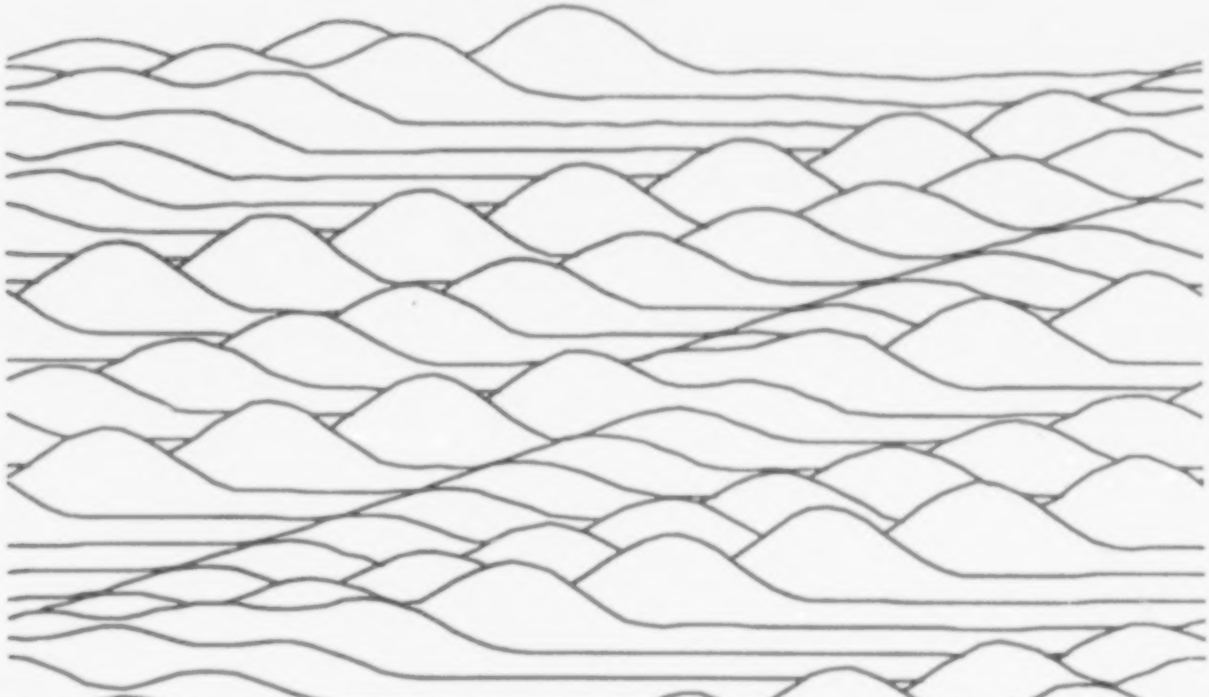


Figure 2. *A space-time plot of three compactons colliding several times in a periodic domain. The compactons experience a phase shift with each collision and create a small ripple.*

The property Kruskal and Zabusky discovered is not unique to the KdV equation. Many other nonlinear equations have the same property: In effect, their traveling-wave solutions (which are also called solitons) are impervious to any kind of disturbance. That makes solitons good candidates for carrying information. If, for example, light can be made to propagate along a fiber-optic cable in accordance with a KdV-type equation, then pulses of digital information can be sent as solitary waves, which travel long distances without distortion.

One drawback of KdV-type solitons, however, is that they aren't truly solitary. Each such "classical" soliton tapers off, both fore and aft, with an infinitely long "tail." As a result, two waves start interacting before their main parts meet (technically speaking, two waves are *always* interacting, but because the tails taper off exponentially, the interaction is weak until the waves are suitably close together). Thus to keep soliton-borne information from getting garbled, the carrier waves have to keep their distance.

That's where Rosenau and Hyman's compact solitons—or "compactons," as the two mathematicians call them—could have an advantage. These new waves are tailless: They vanish abruptly at the endpoints of a well-defined interval. As a result, two compactons cannot interfere with one another until they overlap. In theory, at least, a string of identical compactons could race along an information superhighway like so many manic tailgaters at rush hour—except that on this highway, everyone adheres strictly to the speed limit.

Whether compactons actually have a future on the fiber-optic highway remains to be seen. For now, Rosenau, Hyman, and their colleagues are interested mainly in the light compactons shed on the theory of solitons. One key insight regards the role of a condition known as integrability. The KdV and other classical soliton equations are all integrable. This means, roughly, that their solutions satisfy infinitely many "conservation laws," much as physical systems obey laws such as conservation of energy and momentum. Integrability helps explain solitons' extraordinary stability: The conservation laws constrain the waves so rigidly that they can hardly fall apart.

Compacton equations, however, are not integrable: they satisfy only a handful of conservation laws. Hyman didn't expect much to happen when they numerically smashed two compactons together, but Rosenau urged him to run the computer experiment. "We took one that's traveling fast and one that's traveling slow, and we banged them into each other," Hyman recalls. That the two waves emerged intact, just like ordinary, integrable solitons "was amazing," Hyman says. These unexpected results indicate that the remarkable stability of solitary waves lies deeper than mere integrability.

Still more surprising is a brand new feature, not seen in classical solitons: When two compactons meet, interact, and separate, they leave behind a wake of tiny ripples (see Figure 3). Rosenau and Hyman almost missed this in their first compacton calculations—they thought they saw only some numerical "noise" in the results, stemming from imprecisions in the computation. (All numerical computations are prone to round-off and other errors: one role of mathematical theory is to study such imprecision precisely). "It was only when we were getting ready to write up the results that we decided to do an extra-high-resolution calculation to get rid of this numerical noise," Hyman explains. "When it didn't go away, we started focusing in on it."

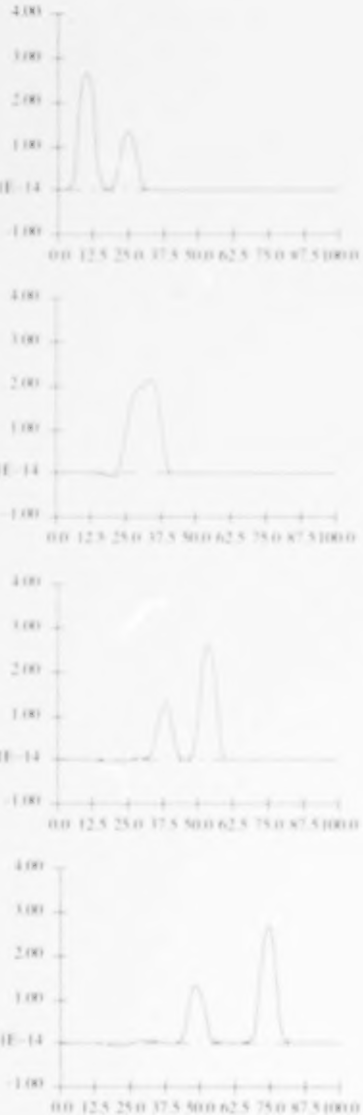


Figure 3. *Ripples result when compactons collide. Here a tall wave overtakes a shorter one as they both move from left to right.*

When two compactons meet, interact, and separate, they leave behind a wake of tiny ripples.

The ripples are a "real mystery," Hyman says. They seem to continue indefinitely, with tinier and tinier ripples arising, a kind of flotsam caused, perhaps, by the compacton equations' lack of integrability. But that's just speculation: There's no proof yet that the ripples don't finally die out, just numerical evidence that smaller and smaller ripples continue to arise. "It's just begging for a solution," notes Hyman. Researchers may no longer chase chance observations on horseback, but plenty of "singular and beautiful" phenomena remain to be found.

A Peek at Peakons

Hyman and colleagues Roberto Camassa and Darryl Holm at Los Alamos National Laboratory have also been looking at yet another new soliton-type equation. Like the compacton equations, this wave equation sports a nonlinear dispersion term, but the new equation also happens to be integrable. This time, the traveling-wave solutions have sharp peaks, hence the name "peakons" (see Figure 4). The researchers believe the peakon equation will provide additional insight into the role of nonlinear dispersion in the theory of solitons. Interestingly, the peakon equation was obtained by simplifying the equations of a global ocean circulation model—the same model that generated the color graphics for the cover of last year's issue of *What's Happening in the Mathematical Sciences*.

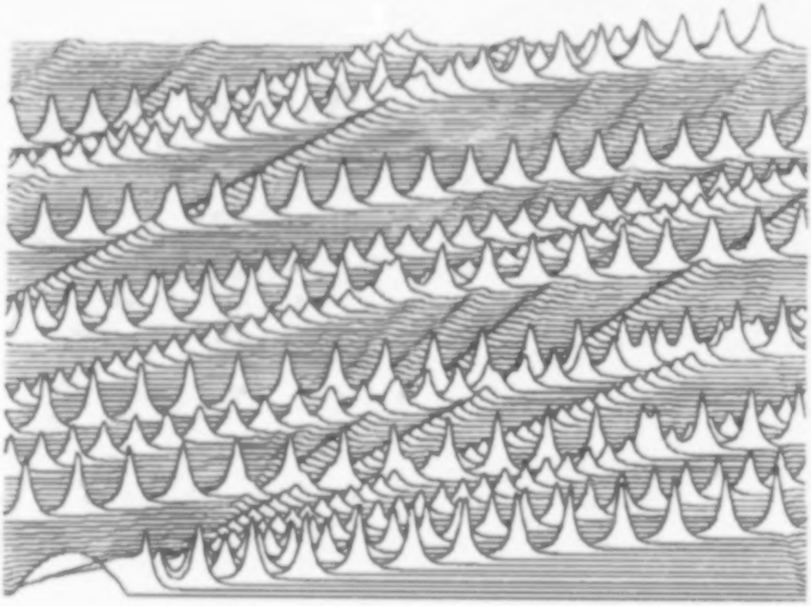


Figure 4 *A space-time plot of peakons—a new kind of solitary wave—generated in this example from an initial parabolic hump (lower left).*

Mathematical Insights for Medical Imaging

Impedance imaging is one of several medical imaging techniques that rely heavily on mathematics.

"First of all, do no harm." Along with his famous oath, the Greek physician Hippocrates left that instruction for his medical heirs. But in order to diagnose disease, modern physicians often find they must perform such invasive procedures as biopsy and angiography. Even X-rays are not without risk. It's a necessary evil: To treat disease, doctors need to see what's happening inside the body. Useful as it is, a stethoscope can't hear cancer cells growing.

Medical researchers are constantly looking for safer, more accurate ways to monitor patients' condition. A team of mathematicians and engineers at Rensselaer Polytechnic Institute (RPI) in Troy, New York, is doing its part to help. Mathematicians David Isaacson and Margaret Cheney, biomedical engineer Jonathan Newell, and their colleagues have developed a new, mathematically-based technology that produces real-time, continuous images of the heart, lungs, and other organs—all without cutting patients open or bombarding them with radiation. They hope their machine, which went into clinical testing at the Albany Medical

Standing, left to right: Jonathan Newell, David Gisser, Gary Saulnier. Seated: Margaret Cheney and David Isaacson. In the left test tank, a piece of conducting pipe and a piece of insulating pipe are visible. These objects are placed in salt water that fills the tanks up to the top of the electrodes. The RPI group tests their impedance imaging system in these tanks. (Photo courtesy of the Rensselaer ACT group.)

BEST COPY AVAILABLE

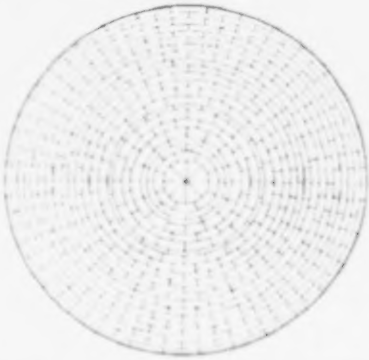


Figure 1 *The 496-grid-cell mesh used to reconstruct images of electrical conductivity distributions. (Reprinted by permission of John Wiley & Sons, Inc., from "NOSER: An algorithm for solving the inverse conductivity problem," M. Cheney, D. Isaacson, J. C. Newell, S. Simske, and J. Goble, International Journal of Imaging Systems and Technology, vol. 2, p. 68, figure 1 (1990). © 1991 John Wiley & Sons, Inc.)*

Center in 1993, will eventually offer physicians a powerful but safe diagnostic tool for such illnesses as heart disease, pulmonary edema, and breast cancer.

The new technology, known as Electrical Impedance Imaging, works by applying tiny electrical currents through electrodes placed on the skin, measuring the corresponding voltage response, and then deducing the distributions of electrical conductivity and permittivity inside the body. (Roughly speaking, conductivity measures how easily charge moves through a medium, while permittivity measures the capacity of a medium to store electrical energy.) Since different parts of the body have different electrical properties, the computed distributions provide an image of the body's tissues and fluids.

Take the lungs, for example. Air is a notoriously poor conductor of electricity. As a result, when healthy lungs fill with air, they show up in an impedance image as regions of low conductivity. By contrast, in a patient suffering from pulmonary edema — a complication often seen following injury, heart attack, or major surgery — the lungs are partially filled with fluid. Because the fluid has high conductivity, the edema appears as an abnormality in an impedance image.

Blood, too, has high conductivity, so impedance imaging also has potential for measuring the amount of blood being pumped by the heart. Measuring cardiac output "is very useful to physicians because it tells them how well the heart is working," explains Newell. "At present, the only reliable ways to measure cardiac output involve passing a catheter through a vein and through the heart, which is a dangerous and expensive procedure."

Impedance imaging is one of several medical imaging techniques that rely heavily on mathematics. The most common is Computed Axial Tomography, or CAT-scan. In essence, a CAT-scan combines X-rays taken from many different directions. Each X-ray measures the density of tissue along a particular line of sight. A computer algorithm based on a mathematical procedure called the Radon transform uses these measurements to reconstruct the actual spatial distribution of densities. Similarly, Magnetic Resonance Imaging, or MRI, constructs images by measuring the body's response to strong magnetic fields.

These techniques all involve solving what are known as "inverse" problems, so called because they ask, in effect, for the opposite of a direct calculation. If, for example, the conductivity distribution in an object is known, then the voltage response to a set of applied currents can be computed directly, much as an algebraic expression such as $2x^2 + 7x - 5$ can be directly evaluated if the value of the variable x is known. On the other hand, the inverse problem — trying to reconstruct the conductivity distribution from a measured set of voltage responses — is like trying to find a value of x for which $2x^2 + 7x - 5$ equals 2, only in a much more complicated mathematical setting.

For impedance imaging, the equations to be solved are derived from Maxwell's equations, a set of partial differential equations that describe all electromagnetic phenomena. Reconstructing an image of the body's interior from measurements on the surface is a considerable challenge, in part because the equations are nonlinear and in part because the reconstruction is highly sensitive to measurement errors. "Conductivity distributions that may be very different may produce data that are very close to each other," notes Isaacson. To cope with that problem, the RPI team has designed a high-precision electrical system for delivering current and measuring voltages, and coupled it with computer algorithms that optimize the system's performance.

The RPI group call their machine ACT III, for Adaptive Current Tomograph.

third generation. It combines delicate engineering with sophisticated mathematical analysis and high-speed computer algorithms to generate precise patterns of current and then reconstruct images from the measured responses. The currents, which are applied through electrodes like those used for electrocardiograms, are well below the level of human perception and considered harmless. That makes the system suitable even for continuous use, as a monitoring device. Whereas a CAT-scan, say, only takes "snapshots" of the body, impedance imaging makes movies, tracking physiological processes in addition to revealing anatomical structure.

That's not to say that impedance imaging will render CAT-scans obsolete. On the contrary, since they measure different properties of tissue, the two technologies complement each other. But impedance imaging offers some special advantages. For one thing, it's relatively inexpensive, in part because of its compact electronics package. It also does not require a specialist to operate or interpret: ACT III or its likely successors could even be used by paramedics on ambulance calls.

Isaacson started studying the mathematics of impedance imaging in the early 1980s. He quickly saw that theory alone was not enough. "I had some ideas about things to do, but I needed some practical experience as to how accurately one can actually measure things, and I wanted to do some experiments," he recalls. He went to Newell, who, while skeptical that impedance imaging could work in practice, helped design an experiment to find out. They decided to see whether Isaacson's electro-mathematics could locate chunks of jello in a tray of saltwater.

"We took a little pan from the local supermarket, filled it with gelatin and salt water, put electrodes around the outside of the pan, pumped some currents in and measured the voltages," says Isaacson. Sure enough, an image appeared, which, though crude, showed roughly where the jello was. That was enough for Newell: "He got very excited about this," Isaacson recalls.

Newell brought in David Gisser, an electrical engineer (now professor emeritus) at RPI, to design and build the electronics. The first system "was crude, but it worked," Isaacson says. The current version incorporates many improvements in both hardware and software. In particular, Cheney notes, Gary Saulnier, an electrical engineer at RPI, and his student Peter Edic have made the system fast enough to work in real time. "A huge number of students have worked on the project at one time or another," Cheney adds. "Ever since I've been associated with the project, there have been somewhere between 10 and 20 students involved at any one time — the number depends mainly on funding. They range from Ph.D. students to undergraduates. We've even had a couple of exceptional high school students."

In experiments with electrodes surrounding a circular tray 30 centimeters in diameter (roughly the size and shape of a human chest), ACT III can reconstruct a reasonable image of a nickel-sized object in the center of the tray — the hardest spot to get a good picture. The RPI group is now also doing experiments with human subjects, including volunteer patients at the Albany Medical Center. To image a 2-dimensional "slice" through the heart and lungs, the researchers place a 32-electrode belt around a person's chest. ACT III then sends a specially designed sequence of current patterns through the electrodes. Voltage measurements taken at the 32 electrodes are fed back to the machine, which uses an algorithm the group calls NOSER (Newton One-Step Error Reconstructor) to produce a circular, 496-grid-cell image (see Figures 1–3). The output is fed to a video monitor, on which the subject can watch — literally live — his or her own lungs filling and emptying

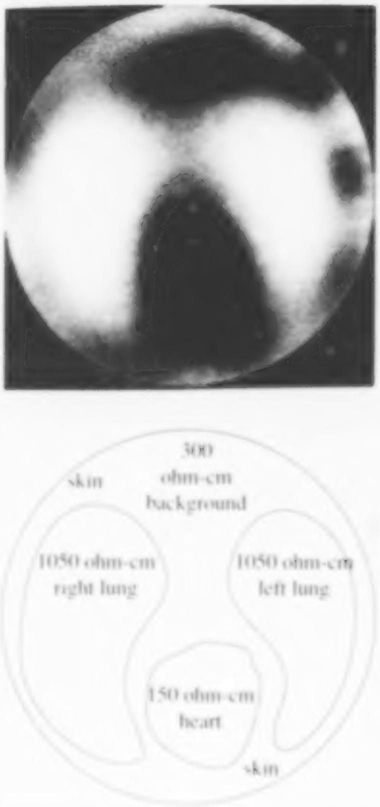


Figure 2 *Impedance image (top) from a test with simulated heart and lungs having specified conductivities (bottom). (Photo courtesy of the Rensselaer ACT group.)*

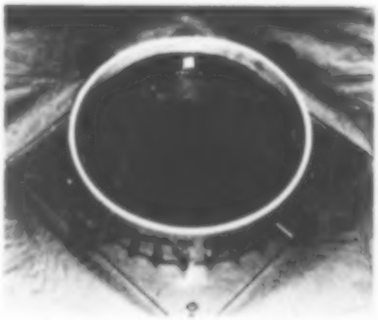


Figure 3 *30-cm test tank with simulated heart and lungs. (Photo courtesy of the Rensselaer ACT group.)*

and blood pumping: Low-conductivity air appears in dark blue, high-conductivity blood — appropriately — in bright red.

Mathematically, each current pattern — say sending current in at just one electrode and taking it out at another — is a vector in 32-dimensional space. More precisely, each pattern is a vector in a 31-dimensional subspace defined by the requirement that the net current applied to the subject must be zero (otherwise the subject's hair would start standing on end). The measured voltage response at the electrodes is also a vector in 31-dimensional space. Roughly speaking, the conductivity distribution is to be found in the matrix that relates the current and voltage vectors. To find that matrix, it's necessary to apply 31 fundamentally different current patterns (in technical terms, the patterns must be "linearly independent").

One key question for the RPI group is which current patterns to use and how to design the electronics to get the best possible signal. In principle, any set that includes 31 linearly independent patterns will do. But that ignores the effect of errors, which can send the linear algebra rattling off into nonsensical solutions.

Isaacson and colleagues have developed the mathematical theory by which ACT III can figure out for itself which current patterns to use.

"It turns out that the best set of patterns to apply depends on what's inside the body," Isaacson explains. For imaging features near the body's surface, patterns that send current in at just one electrode and take it out at an adjacent electrode are optimal. The RPI group, however, uses patterns based on the trigonometric sine and cosine functions. These patterns are provably optimal for distinguishing features deep inside the body. Isaacson and colleagues have developed the mathematical theory by which ACT III can figure out for itself which current patterns to use.

The researchers are also exploring new reconstruction techniques. Cheney has led the way on one promising approach called layer stripping. Conceptually, layer stripping amounts to solving for the conductivity distribution layer by layer, like peeling an onion. The current and voltage measurements, which are made on the outside surface, are used directly to solve for the conductivity of the first layer. From this solution, a set of currents and voltages are computed for the *inside* surface of this layer. These "measurements" are then used to obtain the conductivity distribution of the next layer, and so on. "It's a simple idea," Cheney says. "The nice thing is, it applies to lots of problems."

The RPI researchers are not the only group working on impedance imaging, but Cheney credits Isaacson with having the clearest vision of what can be done. "One of the key things he is able to do is to ask the right questions," she says. "People working on inverse problems usually start by thinking about the reconstruction problem. Figuring out what data one needs in order to do reconstruction often suggests what measurements should be made. But Dave looked at the problem from the point of view of actually building a system, and asked the more fundamental question of how to make measurements containing the maximum amount of information."

The answers could wind up saving lives.

Parlez-vous Wavelets?

Ingrid Daubechies. (Photo courtesy of AT&T Bell Laboratories.)

"Mathematicians are like the French," the German poet Goethe once remarked. "They take whatever you tell them and translate it into their own language — and from then on it is something entirely different."

Goethe's observation is as true now as ever. But times may be changing. In the last ten years, mathematicians and researchers in diverse areas of science, engineering, and even art have discovered and begun to develop a theoretical language they can all understand. This new common language is sparking new collaborations. Many mathematicians are now crossing over into such applied areas as signal processing, medical imaging, and speech synthesis. At the same time, much deep but abstract-sounding mathematics is becoming accessible to researchers in fields from geophysics to electrical engineering.

The new language is wavelet theory. Those who speak it describe wavelets as powerful new tools for analyzing data. Wavelet theory serves as a kind of numerical zoom lens, able to focus tightly on interesting patches of data — but without losing sight of the mathematical forest while attending to the trees, twigs, buds, and grains of pollen.

"Never before in anything on which I've worked have I had contacts with people from so many different fields," says Ingrid Daubechies, a mathematician at AT&T Bell Laboratories and a leading authority on wavelet theory. Because there are so many aspects to the subject, "you have all these ideas brewing together — it's very fertile for everybody concerned." Daubechies adds, "It's a very nice laboratory for showing that applications can have interest for pure mathematics, and vice versa."

Mathematically, wavelets are an offshoot of the theory of Fourier analysis, introduced by the French mathematician Joseph Fourier in his essay *Théorie analytique de la chaleur* (analytic theory of heat), published in 1822. Fourier analysis seeks — with great success — to understand complicated phenomena by breaking them into mathematically simple components. The fundamental idea is to take a function and express it as a sum of trigonometric sine and cosine waves of various frequencies and amplitudes. The familiar and well-understood trigonometric functions are easy to analyze. By combining information about a function's sine and cosine components, properties of the function itself are easily deduced — at least in principle.

Fourier analysis is among mathematics' most widely used theories. It is especially suited to analyzing periodic phenomena, periodicity being the most prominent property of sines and cosines. But even so, the theory has its limitations and its pitfalls. The main problem is that finding detailed information about a function requires looking at a huge number of its infinitely many Fourier components. For example, a transient "blip," obvious in a graph, is impossible to recognize from its effect on a single component. The reason, in essence, is that each sine and cosine wave undulates infinitely in both directions: thus a single wave can't help locate anything. Indeed, the sharper the blip, the more Fourier components are needed to describe it.

Wavelet theory takes a different approach. Instead of working with the infinitely undulating sine and cosine waves, wavelet analysis relies on translations and dilations of a suitably chosen "mother wavelet" that is concentrated in a finite interval. Almost any function can serve as the mother wavelet: this makes wavelet theory more flexible than traditional Fourier analysis. "Daughter" wavelets are formed by translating, or shifting, the mother wavelet by unit steps and by contracting or expanding it by powers of two (see Figure 1). One then expresses other functions as combinations of wavelets, just as Fourier analysis represents functions by combining sines and cosines.

The fact that the mother wavelet is concentrated in a finite interval gives wavelet theory its zoom-in capability: An interesting blip in a function can be analyzed by looking only at those wavelets that overlap with it; finer details are resolved by looking at increasingly contracted copies of the mother wavelet in the vicinity of the blip.

Many of the ideas underlying wavelet theory have been around for decades, but the subject itself got off the ground only recently. The story starts in the early 1980s in France, when wavelets were introduced by geophysicist Jean Morlet and mathematical physicist Alexander Grossmann. In 1985, mathematician Yves Meyer constructed a family of wavelets with two highly desirable mathematical properties, called smoothness and orthogonality. (Interestingly, J. O. Stromberg at the University of Tromso in Norway had constructed such a family several years earlier, but the connection with the nascent theory of wavelets was not realized until after Meyer's work.)

The following year, Meyer and Stephane Mallat gave the subject a solid foundation with a theory of "multiresolution analysis." Then in 1987, Daubechies constructed a family of wavelets that, in addition to being smooth and orthogonal, were identically zero outside a finite interval. Daubechies's construction opened up the field. "Compactly supported" wavelets are now easy to come by, and are among the most commonly used in applications.

And applications are abundant. Wavelets are being tested for use in everything from digital image enhancement — making blurry pictures sharp — to new methods in numerical analysis (itself widely used in scientific computing). "They're a very versatile tool," says Daubechies. Not all the applications will pan out, but many will, and some already have. "There are some very nice success stories," Daubechies adds.

One such story may have far-reaching effects, especially for the next generation of criminals. The Federal Bureau of Investigation has adopted a wavelet-based standard for computerizing its fingerprint files. The FBI has around 200 million fingerprint cards on file, according to Peter Higgins, deputy assistant director of the Bureau's Criminal Justice Information Services division, and 30,000 to 40,000 identification requests pour in every day. At present, the FBI's fingerprint files consume about an acre of office space. The goal, says Higgins, is to digitize the files, store them electronically, and "put [them] in something that would fit in a 20 × 20-foot room."

It sounds easy: after all, entire encyclopedias now fit on a compact disk with room to spare. But that's words. Images are something else. At a resolution of 500 pixels per inch, a standard fingerprint card contains nearly 10 megabytes of data. Transmitting that much information over a modem — something the police would like to be able to do — takes hours at today's transmission rates. For a dozen cards, it's quicker to use Federal Express.

What's needed is some way to compress the data on a fingerprint card without distorting the picture. That's where wavelets come in. By treating the fingerprint image as a two-dimensional function, it's possible to represent it with a combination of wavelets. With a suitably chosen family of wavelets, only a relative handful

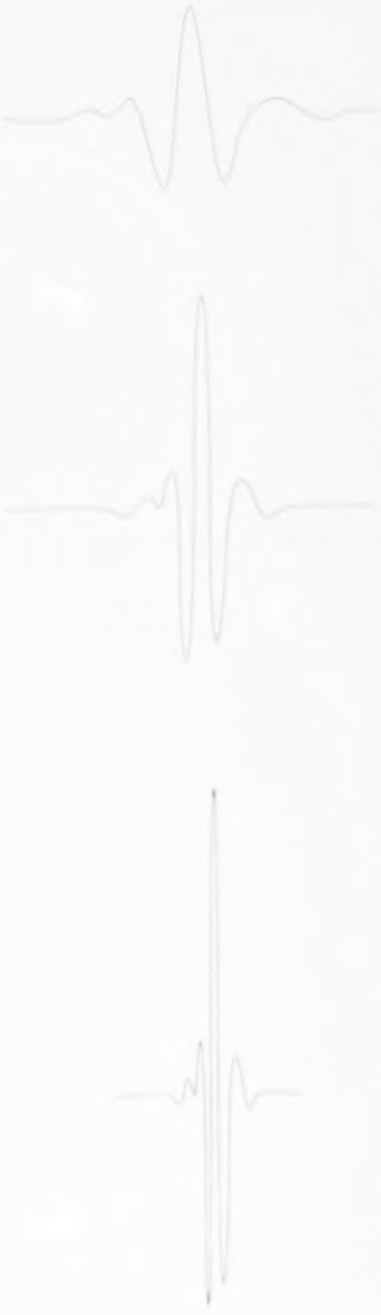


Figure 1. *A "mother" wavelet (top) and two "daughters." (Figure courtesy of Ingrid Daubechies.)*

are needed to represent a fingerprint, and the contribution of each wavelet can be rounded off, or "quantized," which reduces the amount of data that needs to be stored or transmitted.

The wavelet standard for fingerprints was developed by Tom Hopper at the FBI and Jonathan Bradley and Chris Brislawn at Los Alamos National Laboratory. The standard allows many kinds of wavelets to be used—in effect, each electronic fingerprint "card" will include formulas for its particular wavelets, as well as the wavelet representation of the fingerprint itself. So far, one family of wavelets has been approved for use. It compresses fingerprint data by a factor of approximately 20 to 1—reducing 10 megabytes to a much more manageable 500 kilobytes—yet gives images that pass the FBI's automated recognition tests. Indeed, the reconstructed fingerprints look almost exactly like the originals (see Figure 3).

Bradley and Brislawn have also applied wavelet techniques to another kind of data compression: managing the numerical geysers that gush out of supercomputers when running such things as global climate models. "High-performance computers are reaching the point where their ability to churn out data is surpassing our capacity for storing and analyzing it," says Brislawn. In the approach he and Bradley have developed, the computer decomposes the solution (for example, a color-coded map of global ocean temperatures) into wavelets; the user can then control the output by specifying how much detail—that is, how many of the wavelet components—he or she wants to see. One challenge is to figure out how much you can compress the output without sacrificing quantitative capabilities of a model, such as long-term statistical predictions of climatic conditions. Brislawn notes, "This looks like a tough question that we won't be able to answer until we get a better idea of what the models are capable of predicting."

Other researchers are studying the use of wavelets not as post-processing tools, as Bradley and Brislawn are doing, but directly in scientific computation itself. Gregory Beylkin at the University of Colorado has been studying applications

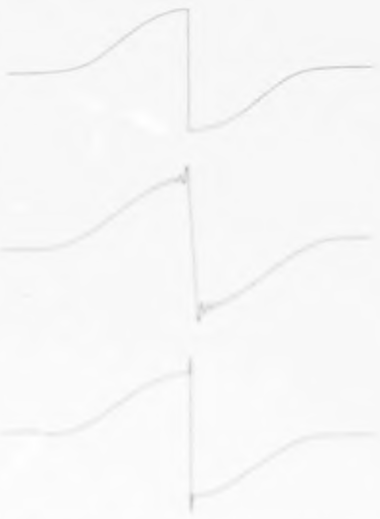


Figure 2 *A Fourier (middle) and wavelet (bottom) reconstruction of a function (top) with a sharp discontinuity. The Fourier reconstruction uses 65 nonzero coefficients, the wavelet reconstruction only 18. (In both cases, the discontinuity causes an overshoot, known as a Gibbs phenomenon, but it is much more localized in the wavelet reconstruction.)*

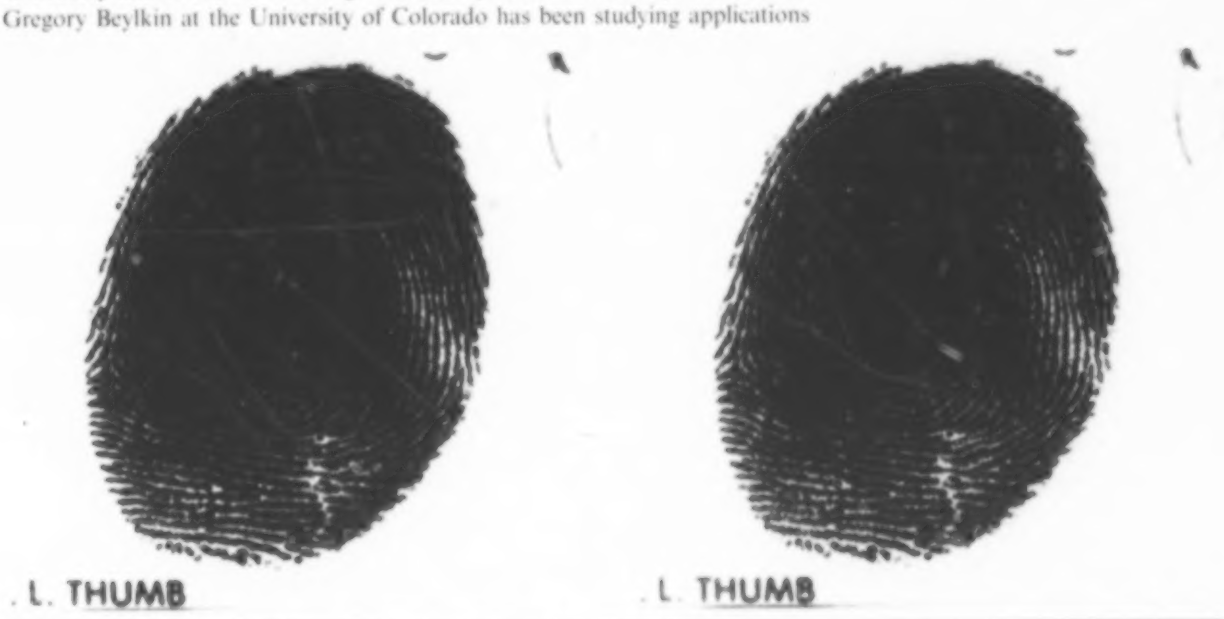


Figure 3 *(Left) Original 768 × 768 8-bit Gray-scale fingerprint image. (Right) Fingerprint image compressed 26.0:1. (Photos courtesy of Chris Brislawn, Los Alamos National Laboratory.)*

of wavelets in numerical analysis. Many problems—such as solving a system of partial differential equations that describes the flow of oil underground—boil down to working with huge matrices, or square arrays of numbers. Such matrices are easier to work with if many of their entries are zero. Beylkin has shown that wavelet analysis can reduce a wide class of matrices to the desired form.

Wavelet theory is making some of the hard-won insights of mathematicians working in the abstract reaches of analysis accessible to researchers in many fields.

Wavelets are especially suited to analyzing sound. Indeed, there's a strong resemblance between wavelets and musical notes. The mother wavelet can be likened to a particular note—say a quarter note at middle C—played at a particular time. Its translates represent the same quarter note at middle C played at other times, while its contractions and expansions are eighth- and half-note C's, played at higher and lower octaves. Ronald Coifman at Yale University and Victor Wickerhauser at Washington University in St. Louis have developed a technique they call adapted waveform analysis, in which a catalog of waveforms is automatically searched for the wavelets best suited to a particular problem. Among the applications is removing noise from recorded sound.

Coifman and his colleagues recently cleaned up an old piano recording of Johannes Brahms playing one of his own Hungarian Dances. Over the years, the recording had acquired several layers of noise. Brahms's live performance was recorded in 1889 on a wax cylinder, which later partially melted. The damaged cylinder was re-recorded on a 78 rpm disk; the version Coifman began with had been recorded from a radio broadcast of the 78. By then the music, competing with pops, hiss, and static, was all but inaudible. Wavelet techniques made it possible to remove enough noise to hear Brahms playing.

Wavelets are also helping researchers clean house in theoretical statistics. "As soon as we were exposed to wavelets, we made the equivalent of about ten years' progress in months," says David Donoho, a Stanford University statistician who has led the way in applying the new theory. Donoho and his colleague Iain Johnstone have developed a "wavelet shrinkage" method for removing numerical noise from data. Their method, which they've shown to be optimal from several technical vantage points, first decomposes data into wavelets and then shrinks each wavelet component according to a rule that eliminates small components altogether.

Donoho expects the insights wavelets supply to set a new agenda for theoretical statistics. Having solved many of the technical problems theorists had long struggled with, "we're in a better position to say what the right questions are for statistical theory to focus on," he says.

Indeed, that may be wavelets' most important legacy. Wavelet theory is making some of the hard-won insights of mathematicians working in the abstract reaches of analysis accessible to researchers in many fields. "Wavelets teach you a way to think about problems so that a lot of ideas in abstract harmonic analysis become natural," Donoho says. The theory of wavelets does more than simply decompose and reconstitute complicated mathematical functions. In Donoho's view, "It's a tool to restructure your thoughts."

Daubechies agrees. "A lot of things are starting to come together," she says. At the same time, she adds, "it's clear that we still need new advances in order to fulfill all the promises that we think are there."

Random Algorithms Leave Little to Chance

It's a common experience: You're walking down an office corridor or a city sidewalk when, without warning, you find yourself face to face with someone in an equal hurry going the other way. You both stop before you collide, and you both step aside—to *your* right. You both smile awkwardly and both step aside again—to your *left*. You both smile again. This time you wait for the other to make a move. You *both* wait for the other. Finally one of you breaks the pattern, and the impasse ends. You both laugh, say "thanks for the dance," and walk away wondering how long you could have both been stuck there.

That scenario generally plays out to comic effect in everyday life. Curiously, something similar occurs in computers—but with effects that are less amusing. When a singleminded program meets the wrong input, the result can be a devastating slowdown. And according to Murphy's Law ("If anything can go wrong, it will"), if there's a data set on which a program runs slowly, then that's the data set the program will be asked to process.

There may be a way out, though. Mathematicians and computer scientists are studying a new approach to programming that avoids the computational gridlock associated with many problems. This new approach relies on a humble but time-honored technique: flipping coins.

The technical term is "randomization," but it boils down to heads and tails. The idea is to insert occasional random decisions into a computation to avoid getting caught up in some unexpected conspiracy between program and data. While random algorithms are susceptible to runs of bad luck, such runs can be made exceedingly improbable. Moreover, that kind of bad luck is independent of the data.

"When you put coin flips into your algorithm, then it doesn't matter how your data is structured," explains Joel Spencer of the Courant Institute of Mathematical Sciences at New York University. "There's no particular kind of data that's bad."

Here's how the idea works in the case of the sidewalk tango. Suppose that you, the program, have a strictly deterministic pattern of responses to the other pedestrian (the data, which in this case is another program). If, say, you always cycle through the responses Left, Right, Wait, then you'll be OK if the data has some other pattern. But if the data happens to be structured the wrong way, then you're stuck forever. On the other hand, if you randomly choose among the possible actions each time, then no matter how the data is structured, it is highly unlikely you'll be blocked for long. Even if the data "wants" to block you, it can't—unless it's somehow clairvoyant, in which case you've got bigger problems. (It's also possible you've stumbled across a mirror.)

Computers, of course, rarely go walking down the sidewalk. A more realistic setting where randomness helps is the task of sorting. Computers are often fed long lists (of names, say, or addresses) to be put into alphabetical or numerical order: it's the kind of "mindless" activity computers excel at. And that's the problem: A machine will gladly spend all day—or all decade—sorting census data, and it will do just that if you don't worry about the efficiency with which it works.

Joel Spencer. (Photo courtesy of Barry Cipra.)

The efficiency of a sorting algorithm (what computer scientists call its "computational complexity") is measured by the number of pairwise comparisons it makes—that is, how often the algorithm compares two objects to see which one comes first. If an algorithm is unlucky (or stupid) it can wind up comparing every pair of items. That's not so bad if you're just trying to put a bridge hand in order. But it's a grim prospect if you've just spilled a thousand alphabetized index cards onto the floor—you could wind up making nearly half a million comparisons. And when the number of items, say on a political party's mailing list, climbs into the millions or tens of millions, the potential worst-case number of comparisons begins to make the national debt look like a pittance.

One popular sorting algorithm is called QuickSort. The basic idea is to choose one item on the list, such as the item currently on top, and then compare everything else with it, forming two piles: those "above" and those "below." The key then is to repeat this procedure with the "above" and "below" piles *separately*—there is no need ever to compare items from different piles. This process, which theorists call "recursive," is guaranteed to work.

On most lists, QuickSort works quite well. More precisely, when averaged over all possible arrangements of a list, the number of comparisons the algorithm makes is proportional to the number of items in the list multiplied by the logarithm of that number. But there are times when QuickSort doesn't work well at all. Ironically, if the list is *already* sorted, then QuickSort does the worst possible thing: It compares every pair of items.

"It's not enough that an algorithm does well on average if there's a patch of problems on which it does very badly, and that patch of problems happens to come up in the real world," says Spencer. "That's exactly the case with QuickSort, because in the real world you *do* sort things that are already sorted."

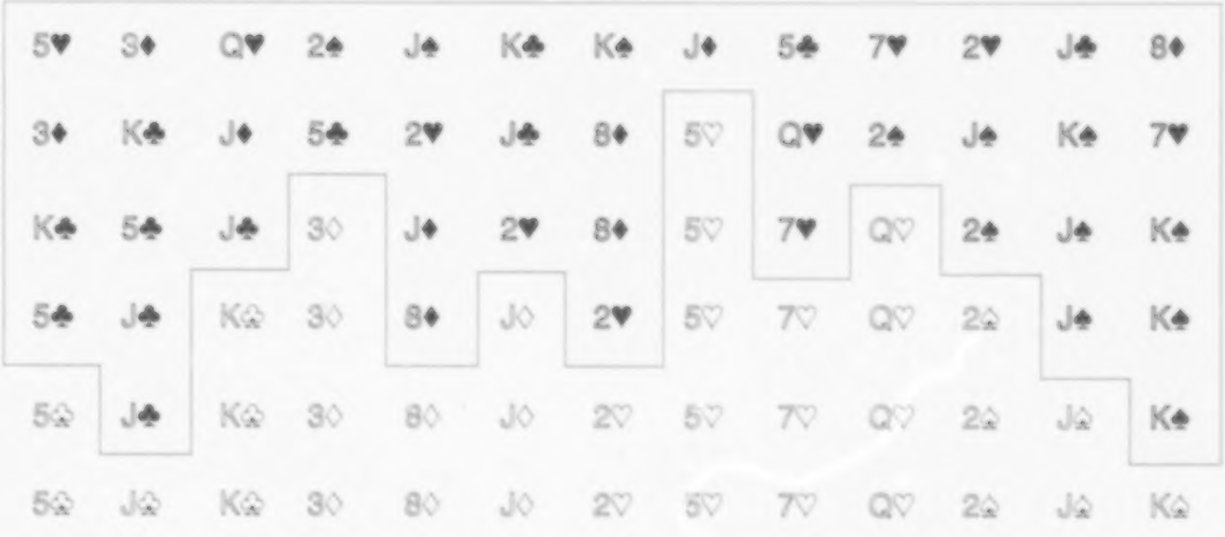

5♥ 3♦ Q♥ 2♠ J♠ K♣ K♠ J♦ 5♣ 7♥ 2♥ J♣ 8♦
3♦ K♣ J♦ 5♣ 2♥ J♣ 8♦ 5♡ Q♥ 2♠ J♠ K♠ 7♥
K♣ 5♣ J♣ 3♢ J♦ 2♥ 8♦ 5♡ 7♥ Q♡ 2♠ J♠ K♠
5♣ J♣ K♧ 3♢ 8♦ J♢ 2♥ 5♡ 7♡ Q♡ 2♤ J♠ K♠
5♧ J♣ K♧ 3♢ 8♢ J♢ 2♡ 5♡ 7♡ Q♡ 2♤ J♤ K♠
5♧ J♧ K♧ 3♢ 8♢ J♢ 2♡ 5♡ 7♡ Q♡ 2♤ J♤ K♤

Figure 1. *QuickSort arranges a bridge hand. In this example, the bridge hand as dealt (top) is rearranged into proper ascending order (bottom) with 30 comparisons. In the worst case, QuickSort makes 78 comparisons to get a hand in order.*

A randomized version of QuickSort solves the problem of "bad" data: Instead of starting with the first item on the list, or any other predetermined item, pick an item *at random*. Most of the time, the two piles will be of roughly the same size. By choosing a random item at each stage, the algorithm will—unless you are exceedingly unlucky—make close to an average-case number of comparisons in total.

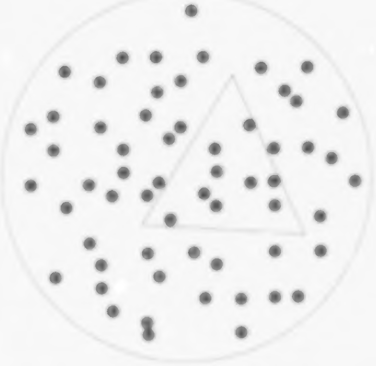


Figure 2 *Nine out of 60 randomly thrown "darts" land inside the triangle, providing an estimate of the triangle's size relative to the circle.*

QuickSort, whether randomized or not, always produces a correct answer—that is, a properly sorted list. What you're gambling on is not the answer, but how long it takes the algorithm to find it. In other applications, an algorithm's run time is guaranteed, but the answer it produces is only approximate, but with a high probability of being very close. One such problem concerns computing (or estimating) the "volume" of an n-dimensional shape. This is not just an arcane mathematical pursuit: the size of higher-dimensional geometric shapes is a central concern in many problems in theoretical physics, chemistry, statistics, and elsewhere.

Theoretical computer scientists have proved that estimating the volume of an n-dimensional shape to a specified accuracy is computationally intractable, if the algorithm used is deterministic. "Intractable" means that the amount of computation increases exponentially with the dimension n, leading to a kind of computational inflation that makes all but the smallest problems too expensive to solve. In general, theorists consider a problem tractable if there is a "polynomial time" algorithm for solving it—that is, an algorithm whose computational demands increase no faster than some power of the size of the problem (in this case, the dimension n). For estimating volume, there is no such algorithm.

No such *deterministic* algorithm, that is.

In 1989, Martin Dyer at the University of Leeds in England and Alan Frieze and Ravi Kannan at Carnegie Mellon University found a random algorithm for estimating volume that broke through the problem's exponential barrier. Their algorithm, which computes the volumes of convex bodies, is based on a method for quickly "getting lost" inside an n-dimensional shape.

The starting point for Dyer, Frieze, and Kannan's algorithm resembles a poorly played game of darts: If you throw darts without aiming, the fraction that hit a particular region of a dartboard is approximately equal to that region's fraction of the dartboard's area (see Figure 2). Curiously enough, an accurate estimate for the region's area requires an *inaccurate* aim.

In n dimensions, the traditional "dartboard" is an n-dimensional "cube," and "darts" are thrown by picking a random number for each of the n coordinates of a point in the cube. But that alone doesn't solve the problem. Estimating volume this way requires a number of darts that grows exponentially with n. The reason is somewhat counter-intuitive: An object can fit snugly into the n-dimensional cube but still occupy just an exponentially tiny portion of the cube's volume. For example, the n-dimensional "sphere" of diameter 1 touching all sides of a unit cube has volume less than $1/2^n$ if $n \geq 12$ (see box next page). Therefore, to have any reasonable chance of estimating the volume of, say, a 100-dimensional sphere, you'd have to throw more darts than there are elementary particles in the universe.

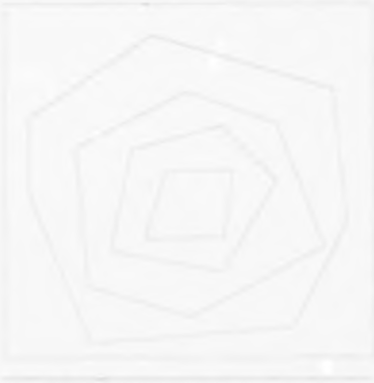


Figure 3 *A nested set of "dartboards" can be used to estimate the size of shapes.*

The three theorists dodge that problem by placing the "target" region inside a nested set of dartboards. The dartboards—really, just convex shapes in n-dimensional space—are crafted so the target occupies a substantial fraction of the smallest, which occupies a substantial fraction of the next smallest, and so on, until the largest dartboard occupies a good bit of the cube (see Figure 3). By randomly

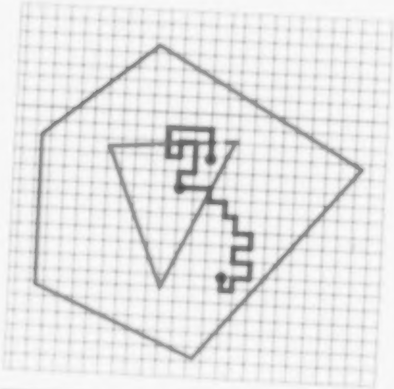


Figure 4 *Two random walks inside a pentagon. Both start at the same point in the triangle, but only one terminates inside the triangle.*

throwing darts at the smallest dartboard, you can estimate the target's volume as a fraction of that board. Likewise, each dartboard's volume as a fraction of the next larger board can be estimated by randomly throwing darts at the larger board. The final result—an estimate for the volume of the target as a fraction of the cube—is obtained by multiplying all these fractions together.

The Incredible Shrinking n-sphere

The area and volume formulas πr^2 and $(4/3)\pi r^3$ are familiar to anyone who has studied circles and spheres. Less familiar, perhaps, is the formula

$$\frac{\pi^{n/2} r^n}{\Gamma((n/2)+1)},$$

which gives the "volume" of an n-dimensional "sphere" of radius r.

The denominator, $\Gamma((n/2)+1)$, takes some explaining. The gamma function, as it's called, is a much-studied special function. It plays important roles throughout mathematics, from geometry to number theory. The gamma function generalizes the factorial function $n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$—itself a central character in combinatorics and probability theory. For computational purposes, the main property of the gamma function is the "recursion relation" $\Gamma(x+1) = x\Gamma(x)$. Thus, for example,

$$\Gamma(4) = 3\cdot\Gamma(3) = 3\cdot 2\cdot\Gamma(2) = 3\cdot 2\cdot 1\cdot\Gamma(1).$$

Likewise,

$$\Gamma\left(\frac{7}{2}\right) = \frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right).$$

To round out the application to the n-sphere, it's enough to know that $\Gamma(1) = 1$ and $\Gamma(1/2) = \pi^{1/2}$. Thus, for example, the volumes of the 4-, 5-, and 6-spheres of diameter 1 (radius 1/2) are $\pi^2/32$, $\pi^2/60$, and $\pi^3/384$, respectively. In general, the volume of the n-sphere gets smaller as n gets larger: The numerator $(\pi/4)^{n/2}$ decreases exponentially, while the denominator $\Gamma((n/2)+1)$ increases. As a result, even though the n-sphere fits snugly inside a cube, the fraction of the cube's volume that it occupies is exponentially small—making it a tiny target if you try playing "darts" on the cube (see main story).

But that strategy only trades one problem for another that seems equally difficult: picking random points inside an arbitrarily shaped region.

Again, if you're not concerned with doing things quickly, there's no problem: All you have to do is use random numbers to generate the coordinates of points in n-dimensional space but discard any points that happen to fall outside the desired region. To keep the computation tractable, however, Dyer, Frieze, and Kannan had to find another strategy—and then prove that it works.

The approach they found involves one of the staples of probability theory: random walks. The new strategy doesn't produce truly random points, but it comes close enough for many purposes, including the n-dimensional volume-estimation problem. The random walks take place on an n-dimensional grid (see Figure 4). Starting from a point that's known to be in the region of interest, the random walker picks one of the $2n$ coordinate directions at random (in three dimensions, for example, she might roll a die to decide whether to go back, forth, left, right, up, or down) and then moves one step in that direction—provided doing so doesn't take her outside the region.

Researchers have known for a long time that a random walker eventually "gets

BEST COPY AVAILABLE

lost" in the sense that after a certain number of steps she has nearly equal probability of being found at any given grid point. The open question was how long it takes to get lost — does the number of steps grow exponentially or polynomially with the dimension n?

"We showed we can get lost in polynomial time," explains Kannan. In other words, even though the number of grid points grows exponentially with the dimension n, the number of random steps it takes to get anywhere on the grid with nearly equal probability grows no faster than some power of n. To prove it, "we needed a fair bit of mathematics," including "various results from differential geometry that had just been proved in the 1980s," Kannan says.

Dyer, Frieze, and Kannan's result applies only to convex regions — and even then, some technical restrictions apply. (It's not hard to see why the method doesn't work in general: If your region is hourglass-shaped, as in Figure 5, then a random walk starting in one compartment may require exponentially many steps to "discover" the other compartment.) The three theorists' original proof showed that the amount of computation required for accurate volume estimates in n dimensions is bounded by n^{23} — a marked improvement over the exponential bounds of deterministic algorithms, but still far from practical. (By contrast, the worst-case behavior of deterministic QuickSort is bounded by n^2.) Subsequent work by a number of researchers, though, has lowered the bound. Most recently, Kannan, László Lovász of Yale University and the Eötvös Loránd University in Budapest, and Miklós Simonovitz of the Hungarian Academy of Science have introduced techniques that lower the bound to n^5 — and, if a certain conjecture is true, down to n^4. The algorithm, which could have a multitude of applications, is now "verging on the practical," Kannan says.

It may be some time before random algorithms become commonplace in computer applications. "There's a real preference among many people in the real world for deterministic algorithms," Spencer acknowledges. But the theory is burgeoning, and the potential is very real. Says Spencer: "There's something that's not coincidental in the effectiveness of randomized algorithms. They're not just a quirk. I think they're really important for computer science."

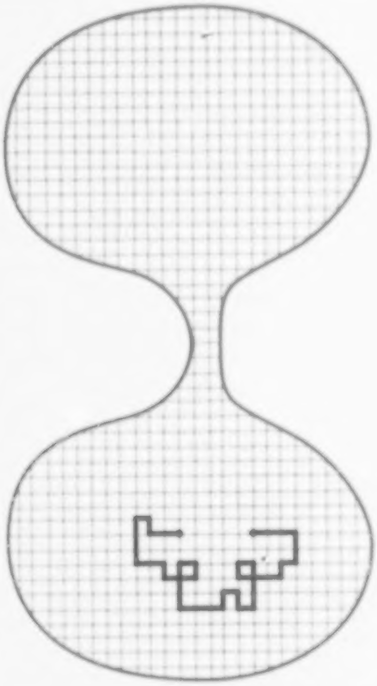


Figure 5. *A random walk in a nonconvex region may take a long time to "explore" the whole region.*

The Guru of Random Algorithms

The guru of random algorithms is a mathematician who has never touched a computer: Paul Erdős, one of the best-known and most colorful mathematicians of the twentieth century. Erdős, who turned 80 in 1993, is a frequent visitor to research centers around the world. He has written hundreds of papers and co-authored many hundreds more. Joel Spencer credits him with invigorating the theory of combinatorics and creating what Spencer calls the probabilistic method, which, while purely mathematical, is what makes random algorithms tick.

One of Erdős's specialties is proving the existence of combinatorial structures without actually constructing them. The probabilistic method, for example, does this by showing that, under the right circumstances, an object picked at random from a certain class of combinatorial objects will have the desired structure with probability greater than zero — and that can happen only if objects with the desired structure exist.

Paul Erdős. (Photo courtesy of Barry Cipra.)

BEST COPY AVAILABLE

The guru of random algorithms is a mathematician who has never touched a computer.

Spencer's favorite example is Erdős's very first: a result in graph theory dating back to 1947. The problem is easiest to state in terms of social engineering: Is it possible to throw a party for, say, *n* guests, at which there are no large groups either of mutual friends or of mutual strangers? (To be precise, by "large" we mean twice the logarithm base 2 of *n*.) Erdős's answer: Yes. Just start with a roomful of mutual strangers, bring each pair together and either introduce them or not, depending on the toss of a coin. By an ingenious proof, Erdős showed that the probability of getting a party with the desired mix is not just greater than zero, it's extremely close to certainty.

That might seem to suggest that Erdős's random introductions could be replaced by some deterministic rule. But so far, no one has found one that works. (The problem, of course, is to find a rule that works for *all* values of *n*.) "No one has even come close to this result by a constructive [algorithm]—and it's been 46 years," says Spencer. The reason, he speculates, is that "when you start to construct things, you're putting structure into them, and this problem seems to demand a *lack* of structure." But who knows? One of Erdős's many protégés might still find a construction that solves the problem.

That's happened with other problems. Indeed, "derandomization" is a hot topic in the theory of random algorithms, Spencer notes. It isn't always possible, and it often makes the inner workings of an algorithm harder to understand, but derandomization offers theoretical insights of its own. At the very least, it gives theorists something to wager over.

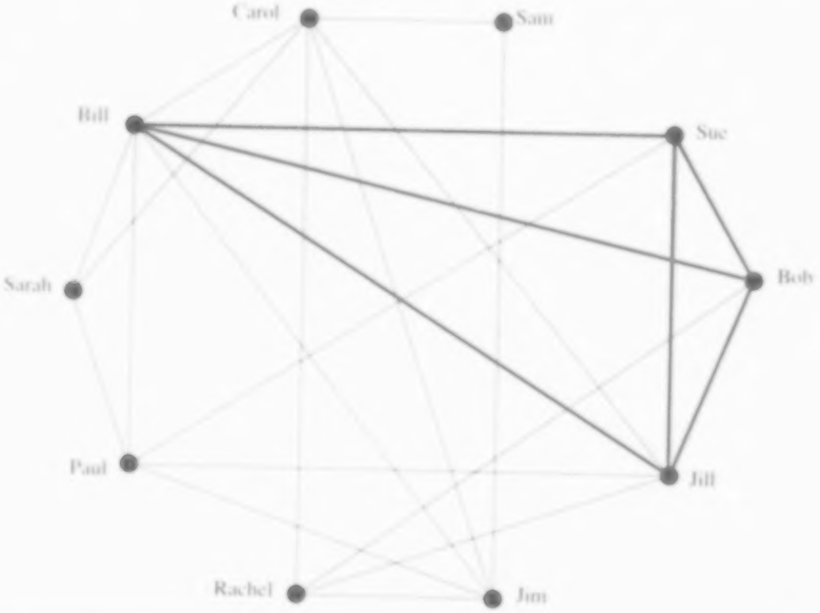


Figure 6. *Ten people at a party. Some know each other (edge), others don't (no edge). Bill, Sue, Bob, and Jill all know each other; Sam, Sue, Rachel, and Sarah are mutual strangers.*

BEST COPY AVAILABLE

Soap Solution

Soap is slippery stuff. So, apparently, is the mathematical theory of soap. Bubbles, that is. When it comes to the geometric properties of soap bubbles, there are more questions than answers. And even when an answer seems firmly in hand, the proof can be as hard to get hold of as — well, as a wet bar of soap.

Chief among the unsolved problems: What shape or shapes will a cluster of soap bubbles assume? It's well known that a *single* bubble minimizes its surface area for the volume it contains by assuming the shape of a sphere. But what happens when *two* bubbles get together? It sounds like a straightforward problem in geometry, with a little bit of calculus thrown in. Surprisingly, the answer is still not known. Or rather, the answer is thought to be known, but so far there is no proof that the answer is correct.

There has been progress, however. For the last several summers, groups of students in a summer Research Experience for Undergraduates (REU) program at Williams College in Williamstown, Massachusetts, have gotten their hands dirty with the theory of soap bubbles. Working with faculty advisor Frank Morgan, an expert in geometric measure theory, the students have taken on — and solved — some subtle problems in the geometry of soap. One group's results appeared in 1993 as a paper in the *Pacific Journal of Mathematics*; other papers are in the pipeline.

The Williams College REU is one of several dozen summer programs around the country offering students a chance to work on open problems in mathematics. More than a hundred students have participated in the Williams program since 1988, working on problems in the theories of knots, numbers, and graphs, along with the geometry of soap bubbles.

Working in the Williams REU "made me look at mathematics differently," says Joel Foisy, who is now a graduate student at Duke University. "Even now it's helping me, because I know that I'm capable of doing original research."

Jeff Brock, now in graduate school at the University of California at Berkeley, agrees. The summer experience was "a definite turning point" in his career. "We were given a tremendous amount of freedom to go about things as we liked," Brock recalls. "When things were going well, I could spend hours at a time thinking about it. And it was really exciting when we actually started getting results."

They had good reason to get excited. In the research leading to their *Pacific Journal* paper, carried out in the summer of 1990, Foisy and Brock, along with Manual Alfaro, Nickelous Hodges, and Jason Zimba, solved the 2-dimensional "double bubble" problem: Given two prescribed areas, find a pair of shapes in the plane whose combined perimeter is as small as possible. In other words, suppose you want to build a pair of corrals of particular sizes (say 1 acre for your sheep and 2 acres for your horse). How should you lay out the corrals to use as little fence as possible?

The suspected answer was the so-called "standard" double bubble (see Figure 1). The two bubbles are separated from the rest of the plane and from each other by circular arcs which meet at angles of 120 degrees (if the two bubbles are of equal size, then the inner arc is a straight line segment). The students showed that every other kind of double bubble has greater perimeter.

Frank Morgan. (Photo courtesy of William Tague, 1989.)

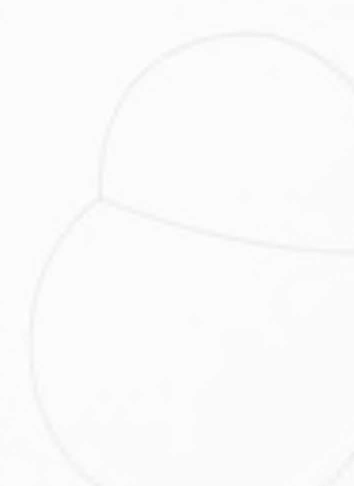


Figure 1. *A standard double bubble. (Figures 1–5 courtesy of Silvio Levy, Geometry Center, Minneapolis, Minnesota.)*

It may seem surprising that this problem hadn't been solved long ago. After all, the single-bubble version, which asserts that the circle is the shortest curve enclosing a given area in the plane, was solved nearly 300 years ago with the advent of the calculus of variations, which treats functions, rather than numbers, as variables. (The problem itself dates back to antiquity. According to legend, when Queen Dido founded Carthage, she "merely" asked for as much land as she could contain within the hide of a bull. She then cut the hide into thin strips, fashioning an enormously long belt, which encompassed a sizable area. Whether Dido actually knew about the area-maximizing property of the circle is unclear, but Carthage was a major power for many years.)

So what held up the double bubble?

For one thing, it wasn't even certain that a solution existed. Conceivably there could be a sequence of increasingly complicated bubble arrangements, each with less perimeter than its predecessor, but not converging to any definite final form. Such problems are perennial in the calculus of variations. For example, among sawtooth curves of fixed length, there is none that minimizes the area beneath the curve (see Figure 2). The area decreases as the teeth get finer, but the "limit" has no teeth at all!

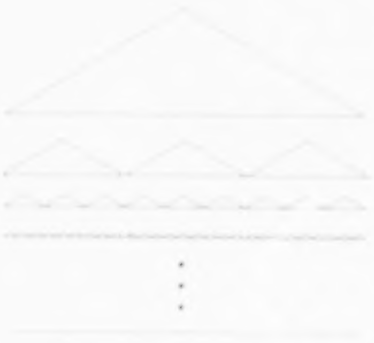


Figure 2 *The limit of sawteeth has no bite at all!*

For bubbles in 3 dimensions (and higher), only in the last 20 years have researchers nailed down the existence part of the theory. In the mid-1970s, Fred Almgren at Princeton University introduced a new geometric definition for soap bubbles and proved that solutions exist to problems of separating specified volumes with minimal surface area. Using Almgren's results, Jean Taylor at Rutgers University proved that these mathematical solutions had the "right" properties: Surfaces meet in threes at angles of exactly 120 degrees, and the seams they form meet four at a time at angles of approximately 109 degrees. These "regularity" properties had been observed in real soap bubbles over 100 years ago, but Taylor was the first to prove that no other behavior is possible.

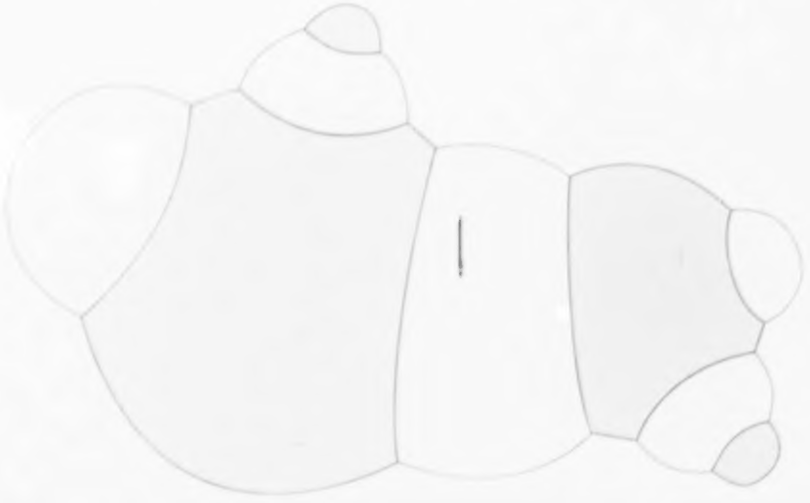


Figure 3 *Could the best double bubble have more than two components? Not in the plane, it can't.*

More recently, Morgan has proved analogous existence and regularity results for 2-dimensional bubbles, both for the plane and for more general (curved) surfaces. Morgan's results gave the students a theoretical basis from which to start.

Even so, there was a lot of work left to be done. "The existence theory admits some very funny things," says Morgan. The main stumbling block is the theoretical possibility that the best way to minimize the total perimeter enclosing several prescribed areas (or the total surface area enclosing several prescribed volumes) is to split each area into several disconnected components. For example, the best "double" bubble might actually be a cluster with nine components, five of which comprise one area and the rest the other (see Figure 3). The existence theory even allows for the possibility that the exterior region has more than one component—that is, there might be "empty chambers" within a perimeter-minimizing bubble cluster (see Figure 4).

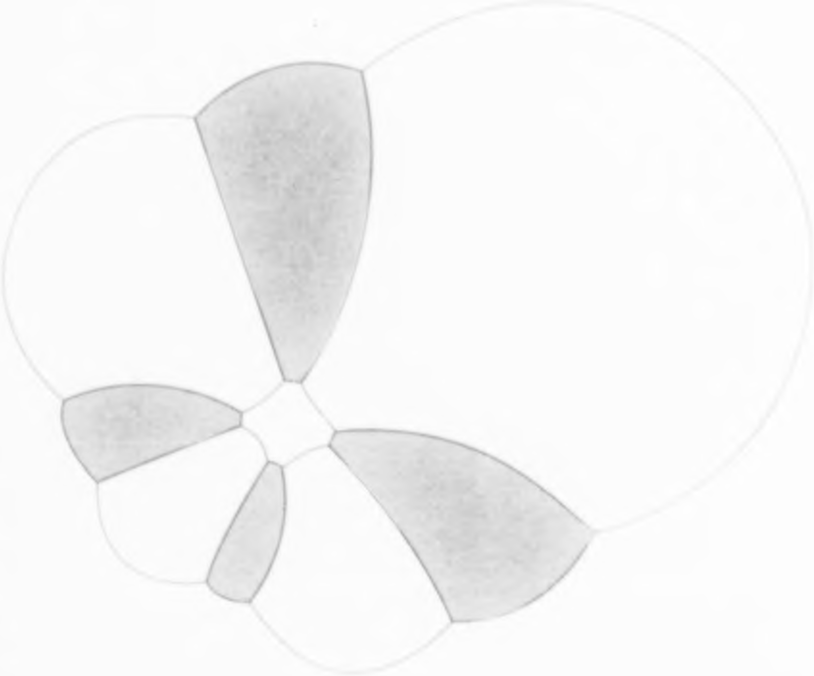


Figure 4 *Could the best double bubble have an empty chamber? Not in the plane, it can't.*

Foisy, Alfaro, Brock, Hodges, and Zimba eliminated those possibilities for the planar double bubble in the summer of 1990. (Brock was an undergraduate at Yale University at the time; the others were students at Williams.) The strategy of their proof had two parts. They first showed that if a perimeter-minimizing double bubble has no empty chambers, then it must be the expected "standard" double bubble; they then showed that empty chambers cannot occur.

In 1992 another group of students took on the 2-dimensional *triple* bubble problem. Chris Cox, Lisa Harrison, Michael Hutchings, Susan Kim, Janette Light, Andrew Mauer, and Meg Tilton showed that, if the perimeter-minimizing

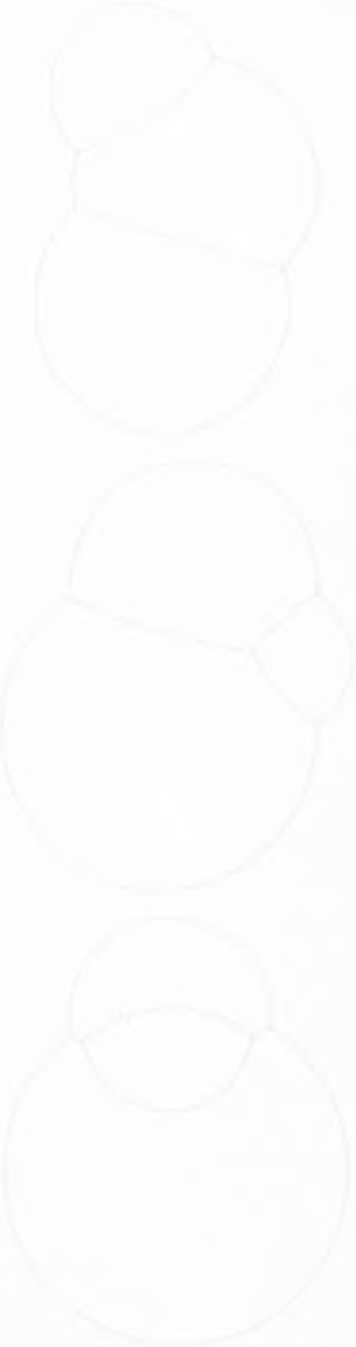


Figure 5 *The standard triple bubble in the plane (center) encloses three specified areas with the least perimeter. Two other types of triple bubbles (top and bottom) do not do as well.*

The last logical steps for the planar triple bubble and the double bubble in space remain to be taken —perhaps by some future group of undergraduates.

solution for three areas is assumed to consist of connected regions, then the "standard" form wins out over two other combinatorial possibilities (see Figure 5). More recently, Hutchings, now a graduate student at Harvard University, has gone back to the double bubble—but up a dimension. It's known that the surface-area-minimizing double bubble is a "surface of revolution" obtained by rotating some plane curve around an axis. The likely answer is the shape that results when the 2-dimensional standard double bubble is spun around its axis of symmetry, but a proof remains elusive. However, Hutchings has shown that the solution, whatever it may look like, has no empty chambers. Moreover, in special cases, such as when the two prescribed volumes are equal, he has proved that the enclosed regions are also connected.

The last logical steps for the planar triple bubble and the double bubble in space remain to be taken—perhaps by some future group of undergraduates. The problems are good ones for students, says Hutchings, because the high-level part of the problem is done, and the rest can be approached with elementary methods: "It just requires some determination and a little ingenuity."

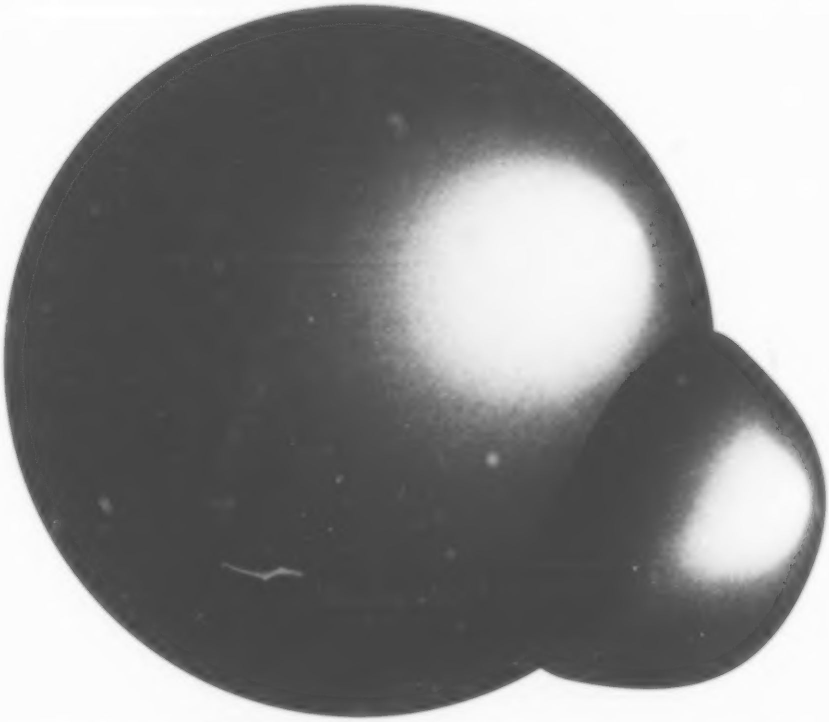


Figure 6 *A 3-dimensional double bubble.*

Straightening Out Nonlinear Codes

Error-correcting code technology is nowadays as common as compact disks: it's what allows your favorite Mozart or Madonna CD to play perfectly even though your cat's been clawing the disk.

Ever since computers started taking over the bulk of the work in data processing and telecommunications, people who use them have worried over a fundamental question: How do you cope when the machine malfunctions?

In the beginning, computers and the programs they ran were simple enough that physical failures—usually the death of a vacuum tube—were readily apparent. But as hardware advanced and programs got more elaborate, the prospect of microscopic flaws that alter how a machine runs or how it handles data became a real concern. With chips getting smaller every year, and computers getting faster, the chance of an occasional error slipping in gets better and better. Even when that chance is one in a billion, a computer running at 25 megaHertz—the speed of last year's laptops, ponderously slow by supercomputing standards—is going to screw up 90 times an hour. What's to be done?

The answer, researchers found, lay in what are known as error-correcting codes. The mathematical theory of these codes, developed over the last 40 years, has enabled computer scientists and engineers to design systems that work reliably at the very edge of their physical capabilities. Error-correcting code technology is nowadays as common as compact disks: it's what allows your favorite Mozart or Madonna CD to play perfectly even though your cat's been clawing the disk. The same technology has been used in deep-space probes, allowing spacecraft such as *Voyager II* to send back sparkling-clear pictures of distant planets while using less power than a refrigerator lightbulb.

Left to right: Roger Hammons, Jr., Patrick Solé, Vijay Kumar, Robert Calderbank, and Neil Sloane. (Photo courtesy of Elizabeth Cruger Arts Photography.)

Error-correcting technology may soon get even better, thanks to some new discoveries in coding theory. Two separate groups of researchers recently found the key to a set of powerful error-correcting codes that have the awkward property of being "nonlinear." Roger Hammons, Jr., at Hughes Network Systems in Germantown, Maryland, and Vijay Kumar at the University of Southern California teamed up with Robert Calderbank and Neil Sloane at AT&T Bell Laboratories

Coding theorists have known for decades that nonlinear codes of a given length can have more code words than their linear counterparts.

in Murray Hill, New Jersey, and Patrick Solé at the Centre National Recherches Scientifique in Sophia Antipolis, France, to show that many nonlinear codes can actually be considered linear — when looked at in the right way.

Their findings, which will appear in the *IEEE Transactions in Information Theory*, "open the gates" to the use of nonlinear codes, says Kumar. Among the promising applications is "sequence design" for digital cellular communications, which will eventually replace the analog technology now used in gadgets such as car phones. Systems based on nonlinear codes could serve many more users with the available bandwidths.

But first of all, what are error-correcting codes, and what does linearity have to do with the subject?

In general, a mathematical code is simply a set of "words," each of which is nothing more than a string of symbols. The most commonly used "alphabet" has just two symbols: 0 and 1. Typically, the words in a code all have the same "length" — that is, the same number of 0s and 1s. (Not all codes work that way. The familiar Morse code, for example, uses a simple *dot* to represent the frequently appearing "word" *e*, but a longer *dot-dot-dash-dot* for the less common *f*.)

The error-correcting capability of a code is based on a notion of "distance" between code words. The distance between two words is simply the number of places in which they differ. If the distance between any two code words is at least 3, then the code can correct single errors. For example, in the code {00000, 11100, 00111, 11011}, the misread word 01100 can be corrected to its "nearest neighbor" 11100, from which it differs in only one digit (it differs from the other code words in two, three, and four places, respectively). Likewise, when the distance between code words is at least 5, the code can correct double errors; triple error-correcting requires distance 7 or more, and so forth.

The code {00000, 11100, 00111, 11011} is also an example of a *linear* code: If you add two code words together using the binary addition rule $1 + 1 = 0$, the result is another code word. For example, $11100 + 11011 = 00111$. Linearity gives a code an algebraic structure that makes decoding messages much easier and makes encoding them a snap. In precise mathematical terms, a linear code is a vector space over the finite field Z_2 — so the full force of linear algebra can be brought to bear.

But linear codes also have their downside. The main problem is that linearity often restricts the number of possible code words, which hinders the code's ability to carry information. If you're not worried about linearity, you can toss in as a new code word any string that maintains the appropriate distance from everything already in the code. But if you insist on linearity, then you have to check these distances not just for the prospective new code word, but also for all its sums with words already in the code.

Coding theorists have known for decades that nonlinear codes of a given length can have more code words than their linear counterparts. Among linear codes of length 16, for example, the best double-error-correcting code has 128 code words. But in 1967, A. W. Nordstrom and John Robinson constructed a nonlinear code containing 256 words (Nordstrom was a high-school student at the time, Robinson an electrical engineer at the University of Iowa). In effect, only 7 digits in each code word of the linear code carry information (the other 9 do the error correcting), whereas in the Nordstrom–Robinson code, 8 digits carry information, an improvement of approximately 14%. Researchers have developed many other

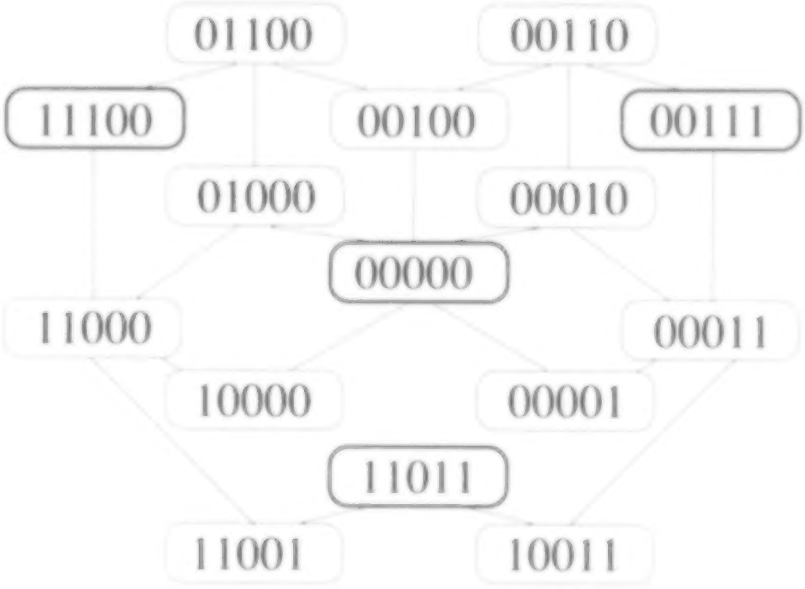


Figure 1 *The code words {00000, 11100, 00111, 11011} and some of their neighbors.*

examples of nonlinear codes, including two families of codes known as Kerdock and Preparata codes, both of which generalize the Nordstrom–Robinson code.

Even so, linear codes have predominated in practice, because their extra structure makes them easier to work with, which translates into faster and more efficient algorithms for encoding and decoding. Nonlinear codes' apparent lack of structure has left them in the dust.

Until now.

The five researchers have discovered a simple trick that turns many familiar nonlinear codes into linear codes — not over Z_2, though, but over Z_4: the algebraic system {0, 1, 2, 3} with the rules $2 + 2 = 1 + 3 = 2 \times 2 = 0$ and $3 + 3 = 2 \times 3 = 2$. More precisely, they have found that many nonlinear codes can be obtained as "images" of codes that are linear over Z_4 using a particularly simple mapping.

The trick is a kind of squaring of the circle (see Figure 2). The system Z_4 is often represented by four points of a compass: 0 and 2 at East and West, 1 and 3 at North and South. Adjacent digits are considered to differ by 1, opposite digits by 2; the distance between code words reflects these differences. For example, the distance between 0000 and 0123 is 4, since the digits differ by 0, 1, 2, and 1, respectively. The new idea is to take a linear code over Z_4 and replace the digits 0, 1, 2, and 3 with 00, 01, 11, and 10, respectively. The result is a nonlinear code over Z_2 that is really just a linear code over Z_4 in disguise!

The surprise is that this trick accounts for essentially all of the nonlinear codes that theorists have studied so far, including the Nordstrom–Robinson, Kerdock, and Preparata codes. It didn't have to work out that way: the trick might have produced only a limited subclass of nonlinear codes — and uninteresting, useless ones, at that. The theory's success hints at deep connections among the various kinds of codes, with more surprises possibly in store.

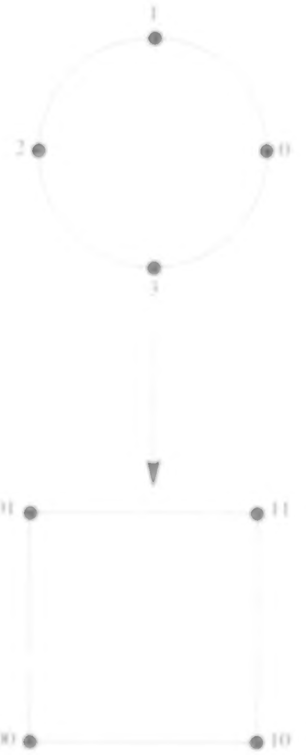


Figure 2 *The key to nonlinear codes: "squaring the circle."*

The theory's success hints at deep connections among the various kinds of codes, with more surprises possibly in store.

The theory has already cleared up a longstanding mystery: the fact that many nonlinear codes seem to come in "dual pairs"—even though "duality" is defined only for linear codes. In particular, each Kerdock code has properties "dual" to those of a Preparata code, while the Nordstrom-Robinson code looks like a "self-dual" code. Researchers had long puzzled over this seeming coincidence. The new results explain it all: the nonlinear codes over Z_2 inherit the appearance of duality from their linear precursors over Z_4, which really are dual.

Hammons and Kumar first discovered a Z_4 connection for the Kerdock and Preparata codes in early 1992. They had been investigating mathematical aspects of a communication technique known as code division multiple access (CDMA), an up-and-coming candidate technology for digital cellular radio. CDMA allows many users simultaneous access to a communication channel by assigning each user a separate code word: the distance between code words prevents users' signals from getting mixed up. With more code words, more people can use the system at once.

Sloane, Calderbank, and Solé made the same discoveries independently later in the year. According to Sloane, David Forney at Motorola Codex and Mitchell Trott at MIT asked him at a conference in October about the possibility of a connection between the Nordstrom-Robinson code and self-dual linear codes over Z_4. Sloane was the right person to ask. He and John Conway at Princeton University had recently completed a study of such codes, so he immediately knew which code would give the connection, if there was one: a self-dual code over Z_4 known as the octacode.

"I went home and, in two minutes thinking about it, it became clear that yes indeed, the octacode was really the same thing as the Nordstrom-Robinson code," Sloane recalls. "I called up Conway and said, 'Look! How could we have missed this? We should have noticed this years ago!' "

Calderbank and Solé contributed several key ideas to Sloane's observation, and the three of them soon had an extensive theory, including efficient algorithms for decoding the Kerdock and Preparata codes. Then Calderbank discovered that Hammons and Kumar had found many of the same results. The two groups agreed to publish their results jointly.

More recently, Sloane and Conway have found Z_4 precursors for a number of single-error-correcting codes that are nonlinear over Z_2, while Calderbank, Kumar, and Tor Helleseth at the University of Bergen, in Norway, have discovered some new codes over Z_4 which are better, by various technical standards, than any of the previously known families over Z_2. Together with Peter Cameron at Queen Mary and Westfield College in London, Bill Kantor at the University of Oregon, and Jaap Seidel at the Technical University of Eindhoven in the Netherlands, Calderbank has also begun investigating a surprising connection between the Z_4-linearity of the Kerdock codes and some seemingly unrelated problems in finite geometry. Nonlinear codes may finally be getting straightened out, but it looks like coding theorists can still count on quite a few twists and turns.

Quite Easily Done

The line between easy mathematical problems and hard ones is finely drawn. Some problems seem to cross back and forth: First they look easy, then they seem hard, and then, when they're finally solved, they look easy again. A recent example is a simple-sounding combinatorial puzzler called the Dinitz problem. First posed in 1978, the Dinitz problem has finally been solved with a surprisingly simple proof, but only after fifteen years during which it seemed a very tough nut to crack.

The story starts in the late 1970s. Jeff Dinitz, then a graduate student at Ohio State University (now a professor at the University of Vermont), was studying properties of combinatorial arrangements known as latin squares. A latin square is an $n \times n$ array of n symbols — say a 5×5 array of stars, squares, circles, diamonds, and triangles — in which no symbol appears more than once in any row or column (see Figure 1). Latin squares are useful, for example, in the design of experiments, to protect against bias. If, say, you want to compare five different herbicides in a corn field, but want to make sure the results aren't affected by variations in soil quality from one side of the field to another, then dividing the field into a 5×5 latin square pattern is an efficient way to design the experiment.

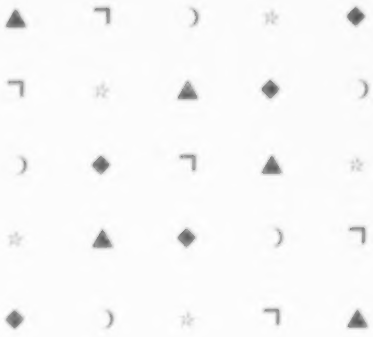


Figure 1. *Each of five symbols appears exactly once in a 5 × 5 latin square.*

Latin squares are easy to come by. Indeed, their number explodes with the size of the square, from two 2×2 squares to twelve 3×3 squares to more than 10^{19} squares of size 8×8. But Dinitz cooked up a variant on the problem of constructing latin squares for which it wasn't clear — until now — that *any* solution could be found.

In an ordinary $n \times n$ latin square, there is only one set of n symbols, and an element from that set must be chosen for each location in the square. In Dinitz's version — called a "partial latin square" — each location is assigned its *own* set of n possible symbols; these sets may vary from location to location. The problem is still to choose a symbol for each location, but now the symbol must come from the set assigned to that location. The goal, however, remains the same: to avoid choosing the same symbol twice in any one row or column.

In Figure 2, a three-element set is assigned to each location in a 3×3 square; the elements in orange constitute a partial latin square. The Dinitz problem asks: Given any assignment of n-element sets of symbols to the n^2 locations in an $n \times n$ array, is it always possible to find a partial latin square? Or to put it negatively, among all the ways to assign n-element sets to the locations of an $n \times n$ array, are there any for which it's impossible to pick an element from each set without picking some symbol twice in the same row or column?

At first glance, the answer seems obvious: Since the problem, in general, uses more than n symbols, it should be easier to satisfy the nonrepetition requirement for a partial latin square than for an ordinary latin square. But that glance overlooks a crucial aspect of the problem: Not every symbol is available at every location. One way to construct an ordinary latin square is to specify where in each row you'll place the first symbol, where the second symbol, and so on; that approach doesn't even make sense for partial latin squares.

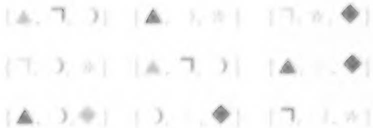


Figure 2. *One symbol (orange) from each three-element set can always be chosen to form a 3 × 3 partial latin square.*

Another telling difference between ordinary and partial latin squares casts further doubt on the "obviousness" of the answer. Ordinary latin squares can always be filled in "row by row." If, say, the first two rows of a 5×5 square have been

Jeff Dinitz

filled in successfully (without doubling up in either row or any column), then the rest of the rows can also be filled in to give a latin square. That means that when you're trying to create a latin square, you'll never paint yourself into a corner—you won't get down to the last row, for example, and find yourself unable to complete the square. With partial latin squares, by contrast, you *can* paint yourself in. For example, if the sets in the first row of a 2 × 2 array are $\{A, B\}$ and $\{B, C\}$, it's natural to choose A and B as the symbols in that row—but then you get in trouble when you see the sets $\{A, C\}$ and $\{B, C\}$ in the next row.

Complications notwithstanding, Dinitz's conjecture—that partial latin squares can always be found—turns out to be true. It just took fifteen years for a proof to be found. In the meantime, the problem served as a kind of drawing card for the theory of combinatorial design and a testing ground for new ideas.

Dinitz's conjecture can be verified directly for 2 × 2 arrays, because there are so few different possibilities. In principle, the conjecture can be checked for arrays of any given size. That's because there are only finitely many cases to check: The total number of distinct symbols for an $n \times n$ array cannot exceed n^3, so the number of cases is less than n^3 to the power n^3 (more precisely, it's at most the n^2 power of $\binom{n^3}{n}$)). But the numbers involved in such a brute-force, case-by-case analysis grow astronomically with n. The 3 × 3 problem is small enough for this approach to be practical, but the 4 × 4 case is already out among the stars.

In 1991, however, Noga Alon and Michael Tarsi at Tel Aviv University in Israel proved a theorem that made it easy to verify (by computer) Dinitz's conjecture for 4 × 4 and 6 × 6 arrays. Their theorem is not specific to Dinitz's problem. It concerns a general problem in graph theory called "list coloring."

In combinatorics, a graph is a set of points (called *vertices*) and a set of lines or curves (called *edges*) connecting them. Many applications of graphs in scheduling or network theory can be interpreted as coloring the edges of a graph, with the stipulation that no two edges of the same color meet at a common vertex. To schedule a college football season, for example, let each team be represented by a vertex, draw an edge connecting teams that are slated to meet, and then color each edge according to the week on which the two teams are to play (say red for week 1, blue for week 2, and so on). The condition that no like-colored edges should meet at a common vertex simply means that no team should be asked to play two games simultaneously.

In a list-coloring problem, each edge in a graph is assigned a prescribed set, or list, of allowed colors. The Dinitz problem can be viewed as a special case of list coloring, for graphs in which each of n "row" vertices is joined to each of n "column" vertices (see Figure 3). Graphs of this type, in which the vertices are separated into two sets and all edges cross from one set to the other, are known as "bipartite" graphs; the particular graph associated with the Dinitz problem is called a complete bipartite graph, because it includes all possible edges between the two sets of vertices. There is a general conjecture regarding how large the palette of possible colors for each edge of a graph must be in order to ensure that a list coloring is possible. Viewed from this angle, the Dinitz problem is just the tip of an immense theoretical iceberg.

Alon and Tarsi's theorem gives a condition which, if satisfied, guarantees the existence of a list coloring from sets of a particular size. Their condition is simple enough to be verified explicitly for the graphs associated with the 4 × 4 and 6 × 6 Dinitz problems. In principle, the condition can be checked for *all* even n.

BEST COPY AVAILABLE

but once again, the amount of computation involved gets quickly out of hand. Furthermore, the condition is *never* satisfied for odd n. (This doesn't mean that the Dinitz conjecture is false for odd n, just that Alon and Tarsi's theorem won't help prove it for those cases.)

Other researchers, notably Roland Häggkvist at the University of Stockholm, had made inroads on the list coloring problem and its relation with the Dinitz conjecture. In late 1992, Jeannette Janssen, then a graduate student at Lehigh University in Bethlehem, Pennsylvania (now a postdoc at Concordia University in Montreal), proved a result that surprised even many of the experts. Janssen showed that Alon and Tarsi's theorem could be used to solve completely a slightly weaker version of Dinitz's problem. Instead of focusing on squares, Janssen looked at *rectangles*—arrays with fewer rows than columns. She showed that in any $r \times n$ array with $r < n$, it's enough to have n symbols (or colors) assigned to each location in order to guarantee that a partial latin rectangle exists.

Janssen's result comes close to the full Dinitz conjecture in two different (but closely related) ways. First, it says that you can always fill in at least the first $n - 1$ rows of a partial latin square (the previous best result guaranteed only two-sevenths of the rows). Second, by starting with an $n \times (n + 1)$ rectangle and then lopping off the last column, Janssen's theorem says that you can always find a

Jeannette Janssen. (Photo courtesy of Cliff Skarstedt.)

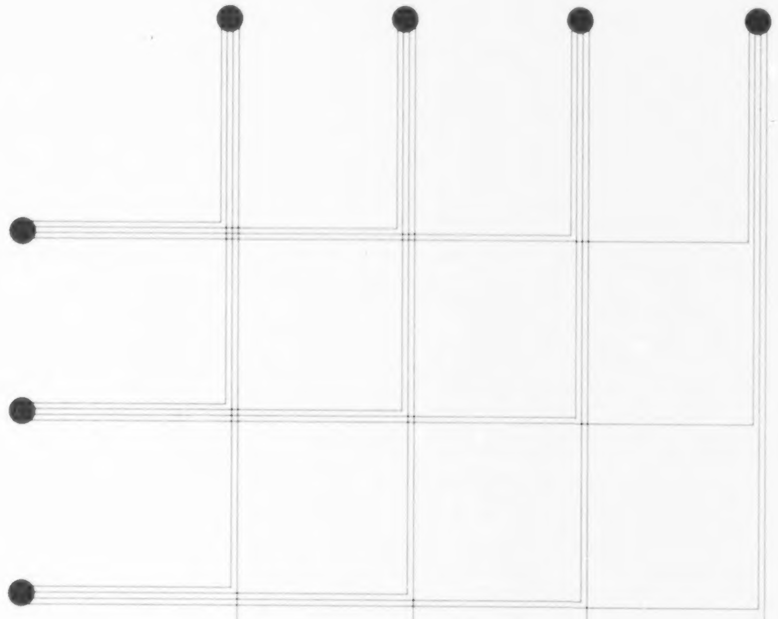


Figure 3. *Each edge in a bipartite graph corresponds to a location in a $n \times n$ array.*

Fred Galvin

partial latin square if $n+1$ symbols have been assigned to each location — again far better than previous results.

Experts in the field lauded Janssen's breakthrough. "It is brilliant," said Herb Wilf of the University of Pennsylvania. "It moves the problem much closer to a resolution than anyone had expected." Other theorists agreed, predicting the full Dinitz problem would be solved soon, perhaps within a year. They were right — but not quite for the reasons they had in mind.

Fred Galvin, a mathematician at the University of Kansas, read Janssen's proof in the *Bulletin of the American Mathematical Society*; this led him back to Alon and Tarsi's paper in the journal *Combinatorica*. A remark in that paper made Galvin realize that one of the ideas in Janssen's work could be parlayed into a proof of the complete Dinitz problem, provided one could prove a certain result about the existence of something called a kernel.

Loosely speaking, a kernel of a graph is a "largest possible" subset of vertices, no two of which are connected by an edge. The precise definition is more technical, but the way kernels are used in Galvin's proof is simple: take any color, say red, identify the set of locations that include red among their allowed colors, find a kernel of that set, and then make red your choice for all the locations in that kernel. The Dinitz problem is solved by repeating this process with other colors until every location has been colored — but this approach wouldn't work, Galvin knew, if some set of locations didn't have a kernel.

"I didn't know much about kernels, so I decided to go to the library and see what's available in the way of kernel existence theorems," Galvin recalls. He found exactly what he needed in the second paper he looked at, a theorem by Frédéric Maffray which appeared in the *Journal of Combinatorial Theory (Series B)* in 1992.

"I was really surprised," Galvin says. "I read and reread [Maffray's paper] several times, thinking maybe I misunderstood one of the definitions." That can happen in a technical tangle of terminology — but not this time. Maffray's theorem was indeed the missing ingredient: the Dinitz problem had been solved.

Galvin circulated a three-page, handwritten account of his findings early this year (1994). He subsequently streamlined the proof to make it self-contained. He is still surprised, almost embarrassed, by the proof's simplicity and the way in which he found it. "None of the ideas in the proof originated with me," he says. "All I did was put together a couple of things that were already in the literature."

The experts are also surprised. "The proof is just amazing," says Jeff Kahn, an expert on combinatorics at Rutgers University. Adds Janssen: "Nobody thought that if there would be a proof, it would fit on three pages."

In fact, Galvin's three-page proof solves the list-coloring problem not just for the complete bipartite graphs associated with the Dinitz problem, but for *all* bipartite graphs. Janssen thinks the proof gives insight into the general list-coloring problem for all graphs. Although Galvin's proof uses none of the elaborate theoretical machinery in Alon and Tarsi's paper or in Janssen's work, the heavy-duty stuff may still be crucial in solving the general problem — the Dinitz problem may have turned out easy to solve because it's a special case, Janssen says. On the other hand, the list-coloring problem may ultimately turn out easy to solve as well, perhaps because it's a special case of some even more general problem. If there's a lesson to be drawn, it's that hard problems need not stay that way.

The Road Least Traveled

Fred Galvin's solution of the Dinitz problem (see main story) not only shows that partial latin squares exist, it also points to an efficient algorithm for finding them. That doesn't always happen: Computer science is rife with problems for which solutions indisputably exist, but for which efficient algorithms to find them are lacking.

In the classic example, known as the Traveling Salesman Problem, a salesman (or woman) starts at the home office, visits a certain set of cities, and returns home. The "cost" (in time, mileage, or money) of traveling between each pair of cities is known: the objective is to complete the calls at the least possible total cost.

The Traveling Salesman Problem is an example of a "combinatorial optimization" problem—"combinatorial" because it deals with ways of arranging things, and "optimization" because it asks for the best arrangement. Like the Dinitz problem, the Traveling Salesman Problem can be phrased in graph-theoretic terms: The vertices of the graph are the cities, and the edges are the roads connecting them.

The problem and its variants have a foot in the door of many applications in which resources need to be routed. The manufacture of printed circuit boards presents one example. In order to connect conductors on different layers of a printed circuit board, it's necessary to drill holes—as many as several thousand, nowadays. The job is best done by a robot, which never gets bored or takes coffee breaks. But even a robot can waste time. The drilling robot must pick up the right size drill bit, move from hole to hole, and then return the bit (perhaps to exchange it for one of a different size). Moving the drill about is necessary, but unproductive and time-consuming; ideally, the drill will move as little as possible.

In principle, solving the Traveling Salesman Problem is easy: If the salesman has n cities to visit (including the home office), then only $(n-1)!/2$ different routes are possible, so it's just a matter of checking to see which one is shortest (or cheapest). The catch, of course, is in that "only." The number of possible routes grows exponentially with the number of cities, making a brute-force approach impractical for any problem with more than a handful of cities.

Computer scientists draw the line at programs whose run time grows exponentially with the size of the problem. They much prefer "polynomial-time" algorithms: programs whose run time grows no faster than some power of the problem size (see "Random Algorithms Leave Little to Chance," pages 27–32). But so far no one has found a polynomial-time algorithm for solving the Traveling Salesman Problem. Indeed, the general consensus is that no such algorithm exists; solutions to the Traveling Salesman Problem, it's thought, are inherently hard to find, even though they obviously exist.

That hasn't kept people from looking for better ways to tackle the problem, though. In part because the Traveling Salesman Problem crops up repeatedly in applications, in part to hone techniques that can be used in other combinatorial optimization problems as well, and in part just because the challenge is there, researchers have developed algorithms which, while still exponential, manage to solve some exceptionally large instances of the problem.

David Applegate at AT&T Bell Laboratories, Bob Bixby at Rice University, Vasek Chvatal at Rutgers University, and Bill Cook at Bell Communications Research (Bellcore) in Morristown, New Jersey, have come up with what might be the best approach yet. Their algorithm stems from a method introduced in 1954, when computers—and combinatorial optimization—were just getting off the ground. The basic idea is to convert the original problem into a sequence of linear programming problems: solving them gives an increasing sequence of lower bounds for the cost of the salesman's cheapest route. Each individual linear programming problem is easy to solve; the catch is, it may require solving a huge number of them to get at the final answer.

Computer science is rife with problems for which solutions indisputably exist, but for which efficient algorithms to find them are lacking.

BEST COPY AVAILABLE

To avoid getting lost in endless computation, Applegate and colleagues have added a "branch and bound" technique. Their algorithm periodically picks a pair of cities and divides the search for an optimal route into two branches: routes that visit the two chosen cities consecutively, and routes that don't. The search along a particular branch is curtailed (bounded) if that branch offers nothing better than a route that's already known.

The new method has already seen some success. Applegate and colleagues have used their branch-and-bound technique to solve more than a dozen longstanding "challenge" problems, including one with 3038 "cities" (see Figure 4 (left)). As it happens, their toughest computation to date is, in a sense, already out of date: One of the challenge problems was to find the shortest tour of all 4461 cities in the former East Germany. The branch-and-bound algorithm chased the problem down a total of 2929 branches before coming up with the answer (see Figure 4 (right)).

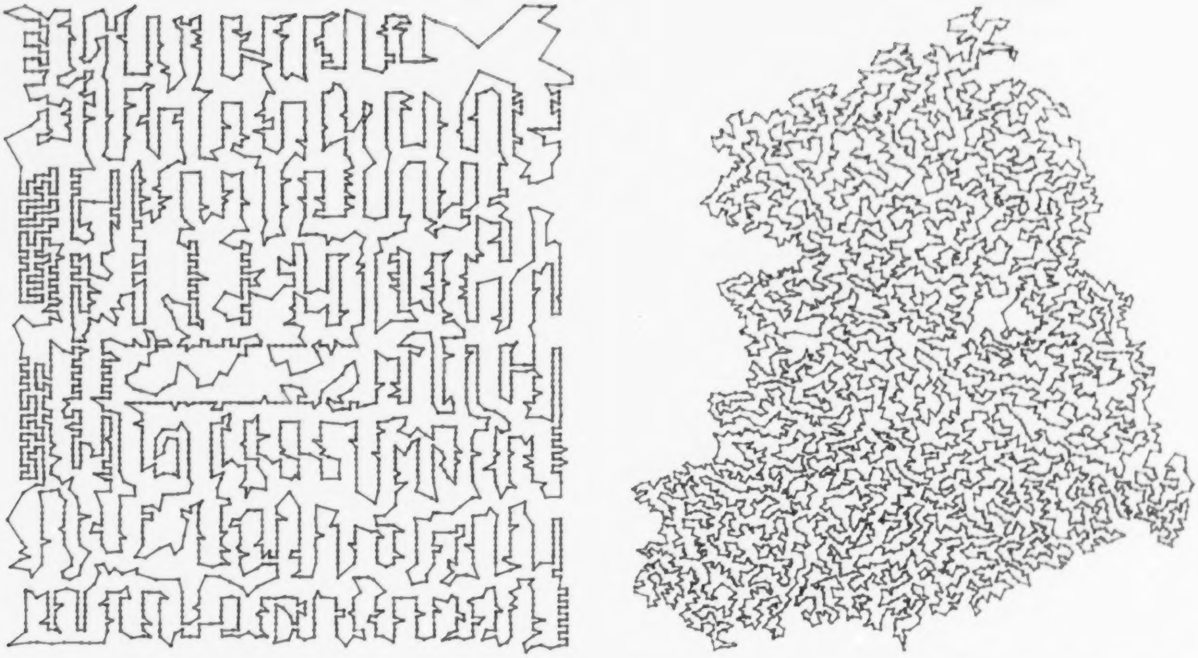


Figure 4. *A traveling salesman's best route around a printed circuit board (left) and the former East Germany (right). (Figures courtesy of Bill Cook, Bellcore.)*

BEST COPY AVAILABLE

(Vector) Field of Dreams

What goes around, comes around, right? Not necessarily. In fact, in the realm of 3-dimensional topology, what goes around need never come back around. At least that's one way to describe a recent result of Krystyna Kuperberg.

Kuperberg, a mathematician at Auburn University in Auburn, Alabama, has resolved a fortysome-year-old problem known as the Seifert conjecture. Dating back to a paper by Herbert Seifert in 1950, the Seifert conjecture concerns the topological properties of a 3-dimensional space, or manifold, known as the 3-sphere. A direct generalization of the ordinary circle and sphere (see box), the 3-sphere is, topologically, the simplest 3-dimensional manifold. But even so, many of its properties remain shrouded in mystery.

The Seifert conjecture, says John Franks, a mathematician at Northwestern University, "was the kind of question that we thought we should be able to answer—and we couldn't." Until now.

In technical terms, the Seifert conjecture asserts that every smooth, nonzero vector field on the 3-sphere necessarily has at least one closed orbit. This sounds eminently reasonable. Indeed, in his 1950 paper, Seifert proved that all vector fields of a certain class (namely, distortions of a vector field known as the Hopf

Krystyna Kuperberg.

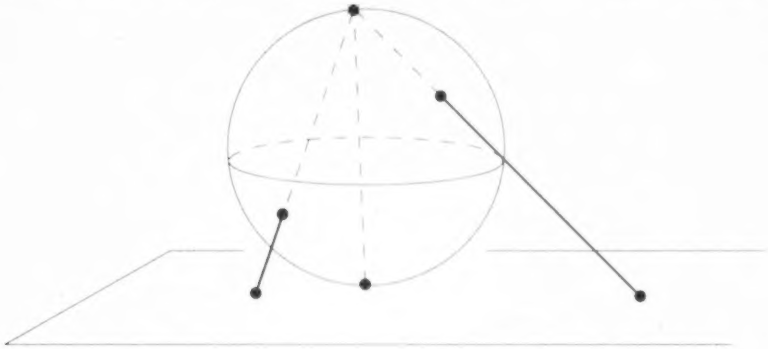


Figure 1. *The "stereographic projection" maps every point in the plane to a point on the 2-sphere by connecting it to the "north pole," which can be thought of as corresponding to a "point at infinity" in the plane.*

Getting 'Round in n Dimensions

The circle, known to topologists as the 1-sphere, or S^1, is the curve defined by the equation $x^2 + y^2 = 1$ in the (x, y)-plane. Likewise, the "2-sphere" S^2 is the surface defined by the equation $x^2 + y^2 + z^2 = 1$ in 3-dimensional space. The n-sphere is a straightforward generalization: It is the n-dimensional "hypersurface" defined by the equation $x_1^2 + x_2^2 + \cdots + x_{n+1}^2 = 1$ in coordinates $x_1, x_2, \ldots, x_{n+1}$. Topologically, the n-sphere is a compact version of n-dimensional Euclidean space with an extra "point at infinity." Figure 1 shows how the (x, y)-plane can be mapped onto the 2-sphere in 3-dimensional space. The corresponding picture in 4-dimensional space is left to the reader's imagination.

BEST COPY AVAILABLE

fibration) do have closed orbits. But even so, mathematicians soon came to doubt the general conjecture. Their doubts were well founded: The Seifert conjecture is false.

By adding an ingenious new twist to some old ideas, Kuperberg has constructed smooth vector fields with no closed orbits, thus putting the kibosh on Seifert's conjecture—and not just for the 3-sphere, but for all 3-dimensional manifolds.

Kuperberg's counterexamples could have implications in the theory of dynamical systems, where closed orbits correspond to periodic behavior, such as the regular swing of a pendulum or the predictable variations in a predator-prey relationship. Vector fields crop up constantly in the study of differential equations and mathematical physics. Newton's law for gravitational motion and Maxwell's equations for electromagnetism are just two examples where vector fields play a key mathematical role in describing physical phenomena.

Loosely speaking, a vector field assigns a little arrow to each point of the surface or space on which the field is defined. Arrows attached to different points can point in different directions, and they can have different lengths. The most familiar example of a vector field is wind: At each point on the surface of the earth, the wind can be described by an arrow pointing downwind, with length proportional to the windspeed. (Of course wind also changes in time. A vector field can be thought of as a wind that varies from place to place, but remains constant in time.) Through each point, a vector field determines a trajectory—the path a dust particle would follow if blown by the field's wind. If the dust particle ever gets blown back to where it began, it will endlessly follow the same path over and over: The trajectory is what mathematicians call a closed orbit.

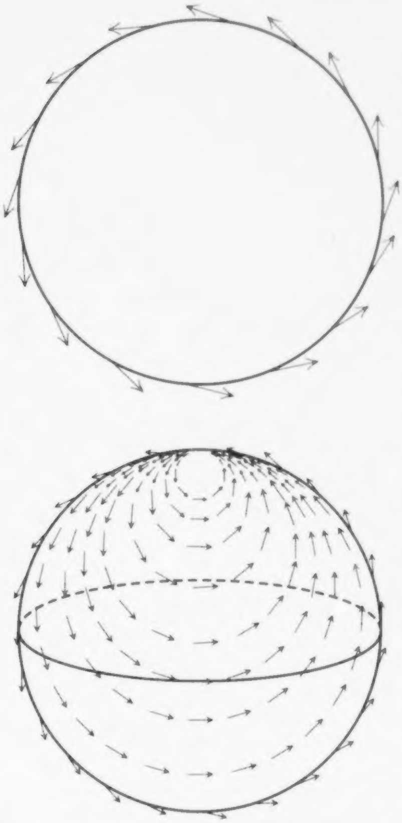


Figure 2. *The 1-sphere (also known as the circle) allows for nonzero vector fields (top), but every vector field on the 2-sphere has a "bald spot."*

There are only two essentially different continuous, nonzero vector fields on the 1-sphere (i.e., the circle): one that points clockwise and one that points counterclockwise. On the 2-sphere, remarkably, there are none at all. Topologists sometimes call this the hairy billiard ball theorem: *You can't comb the hair on a billiard ball, unless it has a bald spot* (see Figure 2). In general, there is a dichotomy between even- and odd-dimensional spheres: Odd-dimensional spheres have continuous, nonzero vector fields, even-dimensional spheres do not.

To restate the dichotomy from a different point of view: Every dynamical system on an even-dimensional sphere has at least one fixed point, whereas dynamical systems on odd-dimensional spheres need not have any fixed points. In effect, Seifert was asking whether dynamical systems on the 3-sphere have the next best property: Do those without fixed points necessarily have closed orbits?

Seifert's original question referred generally to continuous vector fields, not specifically to smooth fields. (A vector field is "smooth" if the lengths and directions of the vectors change not just continuously, but smoothly—in technical terms, if the field is "infinitely differentiable.") But in 1972, Paul Schweitzer at Pontifica University Catolica in Rio de Janeiro, Brazil, produced a once-differentiable counterexample. A decade later, Jenny Harrison at the University of California at Berkeley constructed nonzero, orbitless vector fields that were twice differentiable. (Based on fractals, Harrison's counterexamples could actually be differentiated up to—but not including—three times, given appropriate definitions for fractional differentiation.)

Schweitzer's and Harrison's once- and twice-differentiable counterexamples made Seifert's conjecture seem unlikely to hold in the infinitely differentiable case either. On the other hand, there are plenty of theorems that hold for smooth functions but lose their grip at any lesser level of differentiability. In any event, the constructions seemed stuck at the low-derivative end of things.

Kuperberg's construction breaks that impasse. Expanding on ideas she and Coke Reed, now at the Supercomputing Research Center in Bowie, Maryland, introduced in 1981 to resolve another conjecture about fixed points of dynamical systems, Kuperberg has shown how to modify a smooth vector field so as to break up any closed orbits that might be present. The construction is "very geometric," Kuperberg says. Keeping things smooth was not the hard part of the problem, she explains: "The main difficulty turned out to be not to form additional circular trajectories" in the process of modifying the field.

Kuperberg has shown how to modify a smooth vector field so as to break up any closed orbits that might be present.

The starting point for Kuperberg's counterexample is a smooth vector field with finitely many closed orbits. (It's well known that such fields exist, even though vector fields with infinitely many closed orbits are easier to come by. Schweitzer and Harrison used the same starting point.) The basic tool in Kuperberg's construction is a topological gadget known as a "Wilson plug" a 3-dimensional shape with a vector field that is constant on its boundary and which "traps" at least one trajectory that enters it. The idea is to pick a point on one of the closed orbits, look at a small neighborhood of that point using a coordinate system in which the vector field is constant, and then replace a piece of that neighborhood with the plug, arranging things so that the formerly closed orbit becomes one of the trajectories that enters the plug and gets trapped inside. The trick is to do this without creating any *new* closed orbits.

Kuperberg pulls off the trick in three steps. Curiously, in the first step she constructs a plug that has *two* closed orbits. At this stage the plug looks like a thick washer (see Figure 3). The vector field points straight up on the boundary, but inside the plug, the vectors change direction: Trajectories are deflected counter-

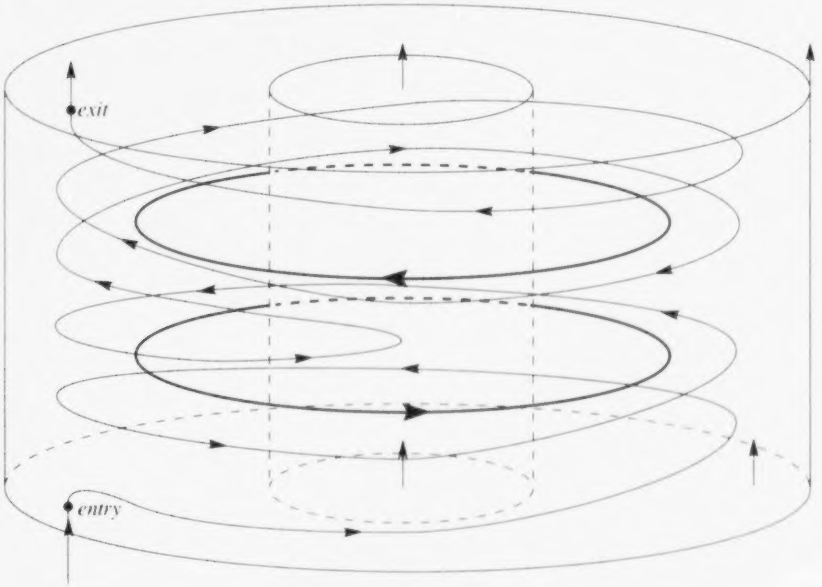


Figure 3. *The first stage in Kuperberg's counterexample to the Seifert conjecture is known as a Wilson plug. Trajectories that enter directly beneath the two circular orbits (dark lines) get trapped inside. (Figure courtesy of Krystyna Kuperberg.)*

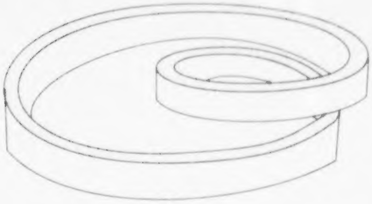


Figure 4. *The second stage in Kuperberg's construction.*

clockwise in the bottom half of the plug, but clockwise in the top half. As a result, any trajectory that makes it all the way through, exits the plug directly above where it enters—as though the field inside the plug were still constant. Trajectories that enter (and exit) near the inner and outer walls are deflected only slightly. The amount of deflection is greater for trajectories that enter away from the walls, until finally, trajectories that enter halfway between the walls pile up on a pair of circles pointing in opposite directions, and thus get trapped inside the plug.

In the second step, Kuperberg refashions the plug to make it look something like a pretzel (see Figure 4). The top and bottom surfaces of the plug no longer lie in a plane, but the walls remain vertical. Most important, there are now places where the inner and outer walls are close together. Finally, in the third step, Kuperberg pinches off two pieces of the plug near the outer wall and stuffs them through the inner wall, giving each piece a twist and skewering it on one of the closed trajectories (see Figure 5). These "self insertions" are the key to her counterexample.

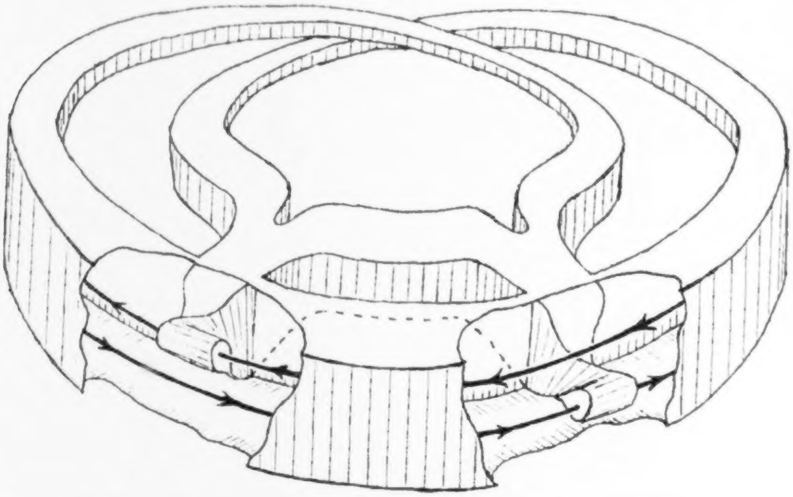


Figure 5. *The final stage in Kuperberg's construction, with cut-away sections to show the self-insertions. (Figure courtesy of Krystyna Kuperberg.)*

"It's the sort of thing that you wouldn't think would be particularly helpful," says Franks. To be sure, the self insertions break up the two closed orbits the plug began with. But it's not at all clear that a slew of new closed orbits aren't created in the process. Indeed, says Kuperberg, "if these self insertions are not chosen the right way, new closed trajectories may form."

To control the trajectories created by the self insertions, Kuperberg relies on what she calls a "radius inequality." Roughly speaking, when a self insertion satisfies this inequality, the resulting plug cannot contain any closed orbits. Instead, all the trapped trajectories spiral endlessly around inside the plug.

These endless trajectories do more than pull the plug on the Seifert conjecture. Kuperberg's construction produces a "minimal set," which John Mather, a dynamical systems theorist at Princeton University, suspects may be of an entirely new kind. Minimal sets are basic components of a dynamical system. Roughly speaking, a set is minimal if the dynamics on the whole set can be generated from the dynamics on any piece of it. In particular, closed orbits are minimal sets, as are fixed points. Other kinds of minimal sets are known, says Mather, but "an overall picture of what minimal sets can be is just lacking." By adding to the list of known examples, the minimal set contained in Kuperberg's plug could help theorists better understand the range of things that can happen in dynamical systems.

Kuperberg's construction also lowers the barrier to proving something much stronger, namely that the 3-sphere itself is minimal for the dynamical system associated with some vector field. Self-minimal manifolds are not that hard to find: the simplest example occurs on the torus (see box). Had the Seifert conjecture been true, the 3-sphere could not have been self-minimal, because a closed orbit can't generate the dynamics away from itself. At this point there's no hard evidence either way, but if it turns out the 3-sphere is self-minimal, that would do much more than refute the Seifert conjecture. It would darn near turn it inside out.

By adding to the list of known examples, the minimal set contained in Kuperberg's plug could help theorists better understand the range of things that can happen in dynamical systems.

Mathematical Donuts

Picture a chocolate cake donut with colored sprinkles all around it, all lined up to point in the same direction (see Figure 6). The non-caloric, mathematical version of this is known as a constant vector field on a torus. If you start at the outer circumference and follow the field once around, either you'll advance along the circumference by a rational multiple of the circumference or you'll advance by an irrational multiple. In the former case, the trajectory from any point will eventually close up: if the trajectory advances by the rational multiple m/n, it becomes periodic after n times around. But in the latter case, the trajectories never close up. Instead every trajectory winds about the torus forever, eventually coming arbitrarily close to every point on the surface. For such a vector field, the entire torus is a minimal set (see main story).

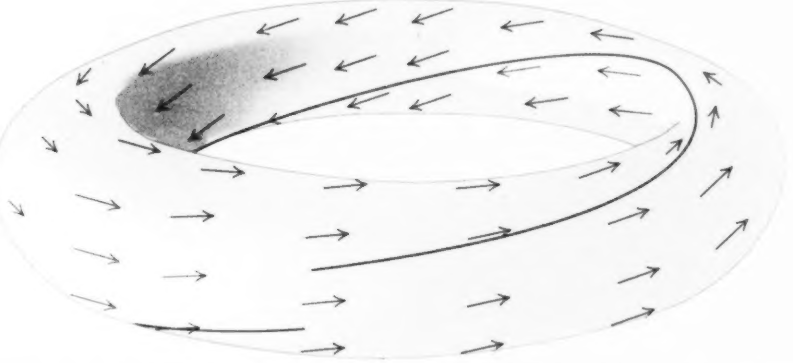


Figure 6. *A vector field on a torus and part of one of its trajectories. The complete trajectory may or may not be closed. (Based on figure courtesy of Frederick Wicklin, Geometry Center, Minneapolis, Minnesota.)*

BEST COPY AVAILABLE

Credits

ADVISORY BOARD

Noga Alon
Tel Aviv University

Randolph E. Bank
University of California, San Diego

Robert Osserman
Mathematical Sciences Research Institute

Carl Pomerance
University of Georgia

Herbert S. Wilf
University of Pennsylvania

About the Author: Barry Cipra, who also did the writing for volume 1 of *What's Happening in the Mathematical Sciences*, is a freelance mathematics writer based in Northfield, Minnesota. He is currently a Contributing Correspondent for *Science* magazine and also writes regularly for *SIAM News*, the newsletter of the Society for Industrial and Applied Mathematics. He received the 1991 Merten M. Hasse Prize from the Mathematical Association of America for an expository article on the Ising model, published in the December 1987 issue of the *American Mathematical Monthly*. His book, *Misteaks…and how to find them before the teacher does…* (a calculus supplement), is published by Academic Press.

About the Editor: Paul Zorn is Associate Professor of Mathematics at St.Olaf College in Northfield, Minnesota. He received the 1987 Carl B. Allendoerfer Award from the Mathematical Association of America for an expository article on the Bieberbach conjecture, published in the June 1986 issue of *Mathematics Magazine*.

Project Administration: Samuel M. Rankin, III, AMS Associate Executive Director

Production Editor: Thomas F. Costa

Production: Ralph E. Youngen, Neil G. Bartholomew, Lori E. Nero, Maxine Wolfson, and Lee Davol.

Design: Peter B. Sykes

The AMS gratefully acknowledges the support of the Alfred P. Sloan Foundation for the publication and distribution of *What's Happening in the Mathematical Sciences*.

About the American Mathematical Society

The American Mathematical Society is a nonprofit organization devoted to research in the mathematical sciences. For more than 100 years, the Society has worked to support the advance of mathematical research, the communication of mathematical ideas, and the improvement of the mathematics profession. In recent years, the Society has increased its attention to mathematics education, public awareness of mathematics, and the connections of mathematics research to its uses.

The AMS is the world's largest mathematical organization, with nearly 30,000 members. As one of the world's major publishers of mathematical literature, the Society produces a wide range of book series, journals, monographs, and videotapes, as well as the authoritative reference, *Mathematical Reviews*. The Society is a world leader in the use of computer technology in publishing and is involved in the development of electronic means of information delivery. Another primary activity of the AMS is organizing meetings and conferences. In addition to an annual winter meeting, the Society organizes bi-annual summer meetings, as well as numerous smaller meetings during the academic year and workshops, symposia, and institutes during the summer. Other major Society activities are employment services, collection of data about the mathematical community, and advocacy for the discipline and the profession.

The main headquarters of the Society, located in Providence, Rhode Island, employs nearly 200 people and contains a large computer system, a full publication facility, and a warehouse. Approximately eighty people are employed at the *Mathematical Reviews* office in Ann Arbor, Michigan. The AMS also has an office in Washington, DC, in order to enhance the Society's public awareness efforts and its linkages with federal science policy.

To order *What's Happening in the Mathematical Sciences*:

Send orders with remittances to:
American Mathematical Society
P.O. Box 5904
Boston, MA 02206-5904

Send VISA or MasterCard orders to:
American Mathematical Society
P.O. Box 6248
Providence, RI 02940-6248

Call toll free in the U.S. and Canada 1-800-321-4AMS (321-4267).
Outside the U.S. and Canada call 1-401-455-4000.

To order Volume 1, 1993, please specify **HAPPENING/1WH**.
Price **$7.00**

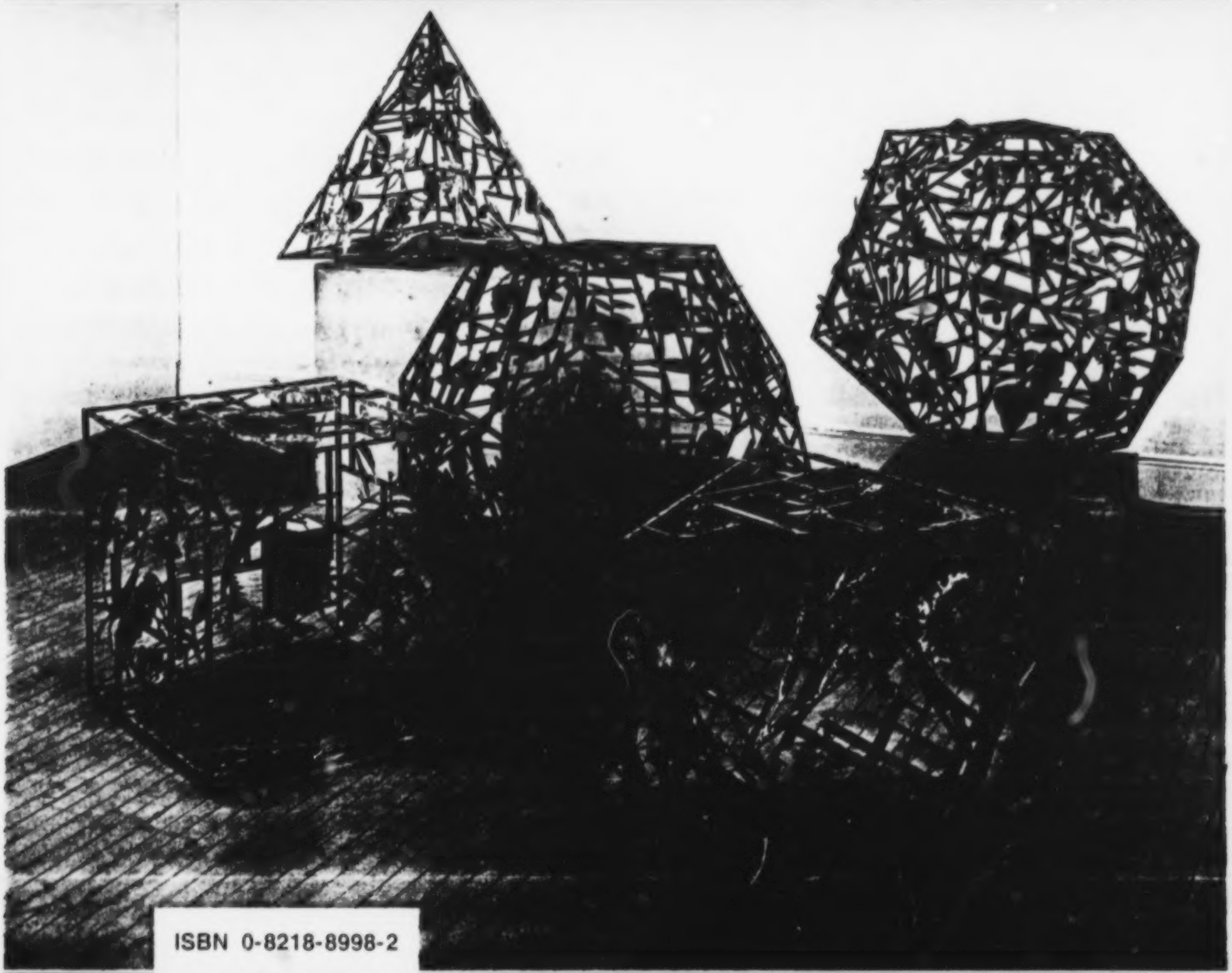
To order Volume 2, 1994, please specify **HAPPENING/2WH**.
Price **$8.00**

To order Volume 3, 1995, please specify **HAPPENING/3WH**.
Price **$8.00**

Call AMS Customer Services for information about bulk order discounts.

Standing orders are available, please contact AMS Customer Services.

All prices effective August 1, 1994. All prices include shipping and handling. For optional air delivery to foreign addresses please add $6.50 per copy. Prices subject to change without notice.



ISBN 0-8218-8998-2

9 780821 889985

BEST COPY AVAILABLE